Analysis of the adaptation mechanism in the type II-A CRISPR-Cas system

DISSERTATION
zur Erlangung des akademischen Grades

Doctor of Philosophy
(Ph.D.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
Shi Pey Wong, Master of Science

Präsidentin
der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

Prof. Dr. Bernhard Grimm

Gutachter/innen

1. Prof. Dr. Dina Grohmann
2. Prof. Dr. Anita Marchfelder
3. Prof. Dr. Emmanuelle Charpentier

Tag der mündlichen Prüfung: 25.02.2019

“It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change.”

– Charles Darwin

To my family.

To everyone who supported and encouraged me.

Table of contents

List of figures

List of tables

List of supplementary figures

List of supplementary tables

Abstract

The RNA-guided adaptive immune system CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated) immunizes prokaryotic cells against mobile genetic elements (MGEs). The CRISPR-Cas immunity operates in three stages, *i.e.* adaptation or spacer acquisition, crRNA biogenesis and interference. During spacer acquisition, a short nucleic acid sequence (prespacer) is acquired from the MGEs, processed and finally integrated into the CRISPR array as a spacer, which serves as genetic memory to defend against the invasion of the cognate MGEs.

Cas1 and Cas2, the key players for spacer acquisition, form an integrase complex that integrates the new spacers into the CRISPR array. The molecular mechanism for the spacer acquisition of the type II-A systems, which encode *cas9*, *cas1*, *cas2*, *csn2* and tracrRNA, is still not fully understood. Therefore, we investigated the architecture of the type II-A CRISPR-Cas protein complex together with the requirement of the different Cas proteins for spacer acquisition.

We verified the acquisition activity of the type II-A systems of *Streptococcus thermophilus* LMD-9 (CRISPR1 and CRISPR3 loci) via spacer acquisition studies by phage challenge. We observed higher acquisition rates in the CRISPR3 locus compared to the CRISPR1 locus. Our plasmid-based spacer acquisition study and concurrent over-expression of Cas proteins, confirmed in addition to Cas1, Cas2 and Csn2 the requirement of Cas9 for spacer acquisition. Crosstalk for spacer acquisition between CRISPR1 and CRISPR3 loci was not observed.

To examine the interactions among the Cas proteins and the potential involvement of non-CRISPR proteins in spacer acquisition, we performed protein-protein interaction studies using yeast two-hybrid and pull-down approaches. These two approaches revealed specific interactions among the Cas proteins, as well as interactions between Cas and DNA repair proteins. We further investigated the interaction between Cas1 and Cas9. The interaction regions of Cas1 with Cas9 were identified by SPOT peptide assay, which demonstrated that the C-terminus of Cas1 interacts with Cas9. Altogether, our study suggests that Cas proteins interact with proteins within and beyond the CRISPR-Cas systems, and it provides a basis for the investigation of the potential roles of DNA repair proteins in the CRISPR-Cas systems and/or *vice versa*.

Zusammenfassung

Das *RNA-guided* adaptive Immunsystem CRISPR (*clustered regularly interspaced short palindromic repeats*)-Cas (CRISPR-*associated*) immunisiert prokaryotische Zellen gegenüber mobilen genetischen Elementen (MGEs). Die CRISPR-Cas Immunität läuft in drei Phasen ab: Adaption oder *spacer*-Gewinnung, Produktion der CRISPR RNS (crRNS) und Interferenz. Bei der Adaption wird eine kurze Nukleinsäurensequenz (*prespacer*) von den MGEs gewonnen, verarbeitet und schließlich als *spacer* in das CRISPR-Array integriert. Dieses dient als genetisches Gedächtnis zur Verteidigung gegen das Eindringen stammverwandter MGEs.

Cas1 und Cas2, die Hauptbestandteile der Adaption, bilden einen Integrase-Komplex, welcher neue *spacer* in das CRISPR-Array integriert. Der molekulare Mechanismus für die Adaptiondes Typ II-A Systems, welches *cas9*, *cas1*, *cas2*, *csn2* und tracrRNA codiert, ist bis heute nicht vollständig verstanden. Daher untersuchten wir den Aufbau des Typ II-A CRISPR-Cas Proteinkomplexes in Kombination mit den Anforderungen der verschiedenen Cas-Proteine für den Adaptionsprozess.

Wir verifizierten die Adaptions-Aktivität von Typ II-A Systemen des *Streptococcus thermophilus* LMD-9 (CRISPR1 und CRISPR3 Loki) anhand von Adaptionsstudien nach Phagen-Infektion. Dabei beobachteten wir höhere Akquisitionsraten im CRISPR3-Lokus im Vergleich zum CRISPR1-Lokus. Unsere Plasmid-basierte Adaptionsstudie bei gleichzeitiger Überexpression von Cas-Proteinen bestätigte die Notwendigkeit von Cas9, zusätzlich zu Cas1, Cas2 und Csn2 bei der Adaption. Ein *crosstalk* von CRISPR1- und CRISPR3-Loki wurde dabei nicht beobachtet.

Um die Interaktionen zwischen den verschiedenen Cas-Proteinen sowie die potenzielle Beteiligung anderer, nicht-CRISPR Proteine bei der Adaption zu untersuchen, führten wir Studien zur Protein-zu-Protein Interaktion durch. Dabei nutzten wir den *two-hybrid,* sowie den *pull-down* Ansatz. Beide Ansätze zeigten sowohl spezifische Interaktionen zwischen den Cas-Proteinen, als auch Interaktionen zwischen Cas-Proteinen sowie DNA-Reparatur Proteinen. Des Weiteren untersuchten wir die Interaktion zwischen Cas1 und Cas9. Die Regionen der Cas1 und Cas9 Interaktion wurden durch *SPOT peptide assay* identifiziert und zeigten eine Interaktion des Cas1 C-terminus mit Cas9. Zusammenfassend weist unsere Studie darauf hin, dass Cas-Proteine sowohl mit Proteinen innerhalb, als auch außerhalb des CRISPR-Cas

Systems interagieren, und bietet somit eine Basis für die Erforschung der möglichen Funktionen von DNA-Reparatur Proteinen in CRISPR-Cas Systemen und *vice versa*.

Abbreviations

Acr	Anti-CRISPR
ABC transporters	ATP-binding cassette transporters
Abi	Abortive infection
AD	Activation domain
BER	Base excision repair
BiFC	Bimolecular fluorescence complementation
bp	Base-pair
BRET	Bioluminescence resonance energy transfer
BREX	Bacteriophage exclusion
BS^3	Bis(sulfosuccinimidy) suberate
Cas	CRISPR-associated
Cascade	CRISPR-associated complex for antiviral defense
CDM	Chemically defined medium
Chi	Crossover hotspot instigator
Cos site	Cohesive end site
CPD	Cysteine Protease Domain
CRISPR or Cr	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
DBD	DNA binding domain
dCas9	Dead Cas9
DSB	Double-strand break
dsDNA	Double-stranded DNA
EMSA	Electrophoretic mobility shift assay
FRET	Fluorescence resonance energy transfer
GAS	Group A *Streptococcus*
His_6-tag	Hexahistidine-tag
HR	Homologous recombination
IHF	Integration host factor
IMAC	Immobilized metal affinity chromatography

IPTG	Isopropylthio-β-d-galactoside
LB	Lysogeny broth
mAU	Mili-Absorbance Units
MBP	Maltose binding protein
MGE	Mobile genetic element
*Mj*Ago	*Methanocaldococcus jannaschii* Argonaute
MMC	Mitomycin C
MOI	Multiplicity of infection
MMR	Mismatch repair
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining
NHS	*N*-hydroxysulfosuccinimide
nt	Nucleotide
NUC lobe	Nuclease lobe
OD	Optical density
PAM	Protospacer adjacent motif
PBS	Predicted biological score
pre-crRNA	Precursor CRISPR RNA
prespacer	Precursor spacer
*Pf*Ago	*Pyrococcus furiosus* Argonaute
Phage λ_{vir}	Virulent variant of phage Lambda
PPI	Protein-protein interaction
REC	Recognition lobe
RM	Restriction modification
rPAM	RNA-protospacer adjacent motif or RNA-PAM
*Rs*Ago	*Rhodobacter sphaeroides* Argonaute
RT	Reverse transcription or reverse transcriptase
RT-PCR	Reverse transcription polymerase chain reaction
SEC	Size-exclusion chromatography
sgRNA	Single-guide RNA
siDNA	Small interfering DNA
Sie	Superinfection exclusion
siRNA	Small interfering RNA
SPR	Surface plasmon resonance

ssDNA	Single-stranded DNA
ssRNA	Single-stranded RNA
THY	Todd-Hewitt broth supplemented with yeast extract
tracrRNA	*trans*-activating crRNA
*Tt*Ago	*Thermus thermophilus* Argonaute
UV	Ultraviolet
WT	Wild type
Y2H	Yeast two-hybrid

1 Introduction

1.1 Prokaryotic immune systems

1.1.1 Phage-host relationships

In nature, phages and their hosts constantly adapt to each other and co-evolve in order to secure their survival. Phages can enhance the fitness of prokaryotes. For example, acquisition of virulence factors encoded in the prophages transforms numerous bacteria, such as *Vibrio cholerae*, *Streptococcus pyogenes*, *Staphylococcus aureus*, *Escherichia coli*, *Clostridium botulinum,* and *Corynebacterium diphtheriae*, to highly virulent strains (Brussow et al., 2004; Keen, 2012). On the other hand, phages can also be harmful to their hosts. For instance, lytic phages are life-threatening to prokaryotes, whereas prophages are an energy burden (e.g. energy for replication, transcription and translation) to their prokaryotic hosts, especially when they do not express any gene that is beneficial for the fitness of the hosts (Vogan and Higgs, 2011). Additionally, random integrations of prophages might disrupt functional genes on the prokaryotic genomes (Vogan and Higgs, 2011). Throughout the bacteria-phage arms race, prokaryotes have developed multi-layer defense systems, which protect them from phage infections. Some of the innate defense systems, *i.e.* defense systems that do not target specific phages based on the record of infections, will be briefly addressed below.

1.1.2 Inhibition of phage adsorption

Phage infection starts with adsorption of the phage on the host cell. Through cell receptor modifications such as mutation, downregulation or masking, bacteria preclude phage adsorption, which is an important step for phage infection (Hyman and Abedon, 2010). However, as a counter-defense to the host receptor modifications, some phages introduce mutations on their receptor binding proteins, which enables them to bind to the modified host receptor (Meyer et al., 2012).

1.1.3 Restriction modification systems

Prokaryotic defense systems may directly neutralize the phage DNAs via restriction modification (RM) systems. A classical RM system employs DNA methyltransferase that modifies only the self-DNA (*i.e.* host DNA), and a restriction endonuclease that digests the non-modified invading DNAs (*i.e.* phage DNAs), thereby protecting the host from the phages. As restriction endonucleases are sequence-specific, phages can evade DNA cleavage by mutations (point mutations) of the restriction sites on the phage DNA, thus preventing recognition by the RM system. (Labrie et al., 2010)

1.1.4 Bacteriophage exclusion system

Bacteriophage exclusion system (BREX) is a six-gene cassette, which contains PglZ phosphatase, PglX methyltransferase, BrxA (a NusB-like RNA-binding anti-termination protein), BrxB (an uncharacterized protein), BrxC (an ATP-binding protein) and BrxL (a Lon-like protease protein). The BREX of *Bacillus cereus* protects against both lytic and temperate phages via DNA methylation to differentiate self-DNA from phage DNA, and subsequent inhibition of the phage DNA replication. In contrast to the RM systems, BREX does not seem to cleave or degrade the phage DNA. (Goldfarb et al., 2015)

1.1.5 Argonaute-based immunity

Prokaryotic Argonautes are speculated for their roles in prokaryotic immunity against foreign genetic elements, however the mechanisms of the prokaryotic Argonautes are still not well understood. Prokaryotic Argonautes demonstrate a broad spectrum of interference mechanisms, including RNA-guided and DNA-guided DNA interference (Hegge et al., 2017). *In vitro* studies showed that the Argonautes from *Thermus thermophilus* (*Tt*Ago), *Pyrococcus furiosus* (*Pf*Ago) and *Methanocaldococcus jannaschii* (*Mj*Ago) are directed by a small interfering DNA (siDNA) guide to cleave DNA targets (Swarts et al., 2014, 2015; Zander et al., 2014). Both *Tt*Ago, and *Mj*Ago are able to cleave double-stranded DNA (dsDNA) in the absence of guides and subsequently use the cleavage DNA fragments as guides for target DNA neutralization (Swarts et al., 2017; Zander et al., 2017). Moreover, *in vitro* studies demonstrated that *Pf*Ago reduces plasmid transformation efficiency (Swarts et al., 2015). On the other hand, a heterologous system in *E. coli* showed that the Argonautes from *Rhodobacter sphaeroides* (*Rs*Ago) binds small interference RNA (siRNA), and the ribonucleoprotein complex reduces the number of

plasmids and plasmid transcripts via a yet to be elucidated mechanism (Olovnikov et al., 2013). SiRNAs sequences predominantly match ribosomal RNA sequences and mRNAs of plasmids, however, the underlying mechanism is still unclear (Olovnikov et al., 2013).

1.1.6 Abortive infection

A distinct defense system, named abortive infection (Abi) aims at defending the whole bacterial population by sacrificing the infected cells. Abi systems generally target essential cellular mechanisms such as replication, transcription and translation. (Labrie et al., 2010))

In the recent decade, the adaptive prokaryotic defense systems known as CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR-associated proteins) systems have been rapidly characterized and harnessed as valuable tools in the fields of biotechnology and genome-editing. CRISPR-Cas systems will be addressed in details in the following sections.

1.2 The adaptive immunity: CRISPR-Cas systems

CRISPR-Cas systems are RNA-guided adaptive immune systems that protect prokaryotes against mobile genetic elements (MGEs), such as bacteriophages and plasmids. CRISPR-Cas systems are detected in around 90% of sequenced archaeal and 50% of bacterial genomes (Grissa et al., 2007; Makarova et al., 2015). The hallmark of the CRISPR-Cas systems is the CRISPR array, which is composed of a succession of repeat sequences interspaced with sequences generally derived from MGEs - termed spacers - that serve as a genetic memory of past infections (Mojica et al., 2005; Pourcel et al., 2005). Upstream of the CRISPR array is an AT-rich leader sequence, which contains the promoter for the transcription of the CRISPR array. Proximal to the CRISPR array is a set of *cas* genes (Jansen et al., 2002) that encodes Cas proteins responsible for the functionality of the CRISPR-Cas systems.

1.2.1 The classification of CRISPR-Cas systems

Based on the presence of the hallmark *cas* genes and the nature of the interference effector complex, CRISPR-Cas systems have been classified into two classes, which are additionally

divided into six types and many subtypes (Makarova et al., 2011, 2013, 2015; Shmakov et al., 2015). Class 1 CRISPR-Cas systems include type I, III and IV, and employ multi-subunit Cas effector complexes for targeting MGEs; whereas class 2 (type II, V and VI) systems rely on a large single Cas protein with multiple domains (Makarova et al., 2015; Shmakov et al., 2015).

1.2.2 The three stages of CRISPR-Cas immunity

The CRISPR immunity is executed predominantly by Cas proteins via three stages: (1) adaptation or spacer acquisition, (2) CRISPR RNA (crRNA) biogenesis and (3) interference (**Figure 1**). Adaptation begins with the detection of the MGEs, from which short nucleic acid sequences (termed as precursor spacers or prespacers (Westra and Brouns, 2012)) are acquired and integrated into the CRISPR array as spacers. The key proteins for spacer acquisition are Cas1 and Cas2 (Datsenko et al., 2012; Heler et al., 2015; Swarts et al., 2012; Wei et al., 2015a; Yosef et al., 2012), which are almost ubiquitous in all the CRISPR-Cas types (Makarova et al., 2011, 2015). During crRNA biogenesis, the CRISPR array is transcribed into a precursor CRISPR RNA (pre-crRNA), which will be processed into mature crRNAs. Each of the crRNAs consists of portions of both repeat and spacer and is bound to either a multi-subunit Cas complex (class 1) or a single-subunit Cas protein (class 2) to form a crRNA-effector complex. Upon the re-invasion of a MGE, crRNA guides the Cas effector complex towards the MGEs to cleave the target sequence (known as protospacer) that is complementary to the spacer portion of the crRNA, *i.e.* a process known as CRISPR interference.

1.2.2.1 Prespacers and protospacers

Prespacers and protospacers are essentially the same nucleic acid sequences on the MGEs. While prespacers are short foreign sequences that will be selected, processed and incorporated into the CRISPR array as spacers, protospacers refer to the short sequences on the re-invaded MGEs that match the existing spacers in the CRISPR array (Westra and Brouns, 2012). While some publications in the literature refer to the precursor of the spacer with the terminology “prespacer”, “protospacer” is more commonly used. In this thesis, the terminology “prespacer” will be used to specify the precursor sequences that will be processed and integrated into the CRISPR array.

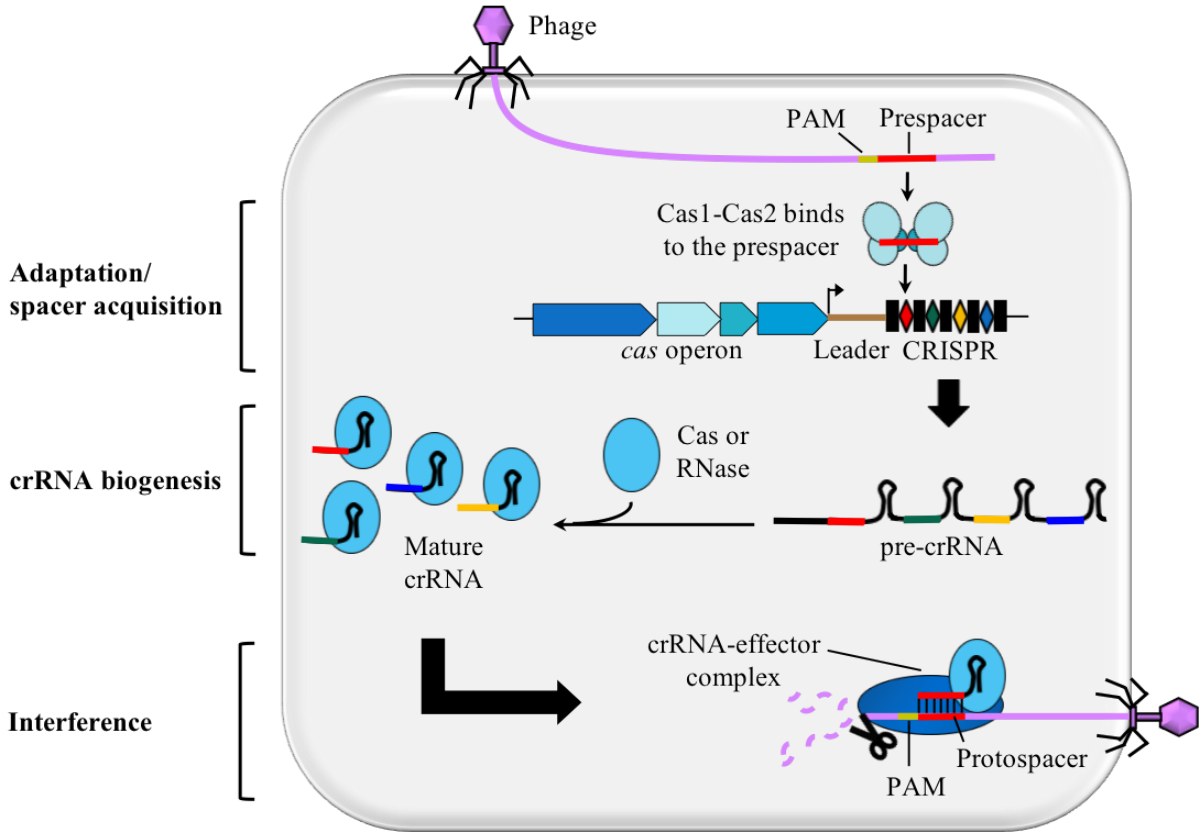


Figure 1. Three stages of CRISPR-Cas immunity.

Upon phage invasion, CRISPR immunity operates in three stages. The first stage is known as adaptation or spacer acquisition, where the adaptation machinery selects a short foreign sequence (prespacer) adjacent to a PAM and integrates it into the leader-end CRISPR array as a spacer. The adaptation machinery involves the Cas1-Cas2 complex; depending on the CRISPR-Cas types, the adaptation machinery may include additional Cas proteins. In the second stage, the CRISPR array is transcribed into a long precursor CRISPR RNA (pre-crRNA), which is further processed into mature crRNAs by either Cas proteins or host RNase(s) during crRNA biogenesis. In the interference stage, a crRNA-effector complex, formed by the binding of a crRNA to either a multi-subunit Cas complex (class 1) or a single Cas protein (class 2), recognizes the cognate phage via PAM identification and sequence complementarity between the crRNA and protospacer. Upon target binding, the Cas nuclease digests the invading DNA. Rectangles, repeats; diamonds, spacers. The colors of the spacers indicate variable MGEs-derived sequences. (Figure adapted from (Hille et al., 2018))

1.2.2.2 Protospacer adjacent motif

CRISPR immunity relies on the insertion of short sequences from the MGEs into the genomic CRISPR loci, which raises the question about how CRISPR-Cas systems distinguish between self-DNA (spacers in the CRISPR array) and non-self DNA (protospacers on the MGEs). In almost all CRISPR-Cas systems (except type III systems), the recognition of a protospacer adjacent motif (PAM), which is a short sequence flanking the protospacer (Deveau et al., 2008;

Mojica et al., 2009), is critical for the MGEs recognition and neutralization. The PAM is present on the MGE but absent on the CRISPR array, and is therefore important for self- versus non-self discrimination during interference to prevent autoimmunity. In view of the importance of the PAM during interference, selecting a prespacer with a canonical PAM during spacer acquisition is also crucial for CRISPR-Cas immunity (Datsenko et al., 2012; Deveau et al., 2008; Mojica et al., 2009; Swarts et al., 2012). In the type III-B system of *Pyrococcus furiosus*, an RNA-PAM (rPAM) located at the 3′-end of the target RNA that is complementary to the crRNA is recognized as a non-self genetic element, which will be cleaved by the interference machinery (Elmore et al., 2016). Unlike type III-B system, type III-A system of *Staphylococcus epidermidis* does not depend on PAM or rPAM for self- versus non-self discrimination (Elmore et al., 2016). Instead, DNA cleavage only occurs when the potential target sequence is complementary to the spacer portion of the crRNA, but not the 5′-repeat handle, the partial repeat sequence on the 5′-end of the crRNA (Marraffini and Sontheimer, 2010).

1.3 The type II-A CRISPR-Cas systems

1.3.1 The subtypes of type II systems

The multi-domain effector Cas9 is a hallmark protein of the type II CRISPR-Cas systems, which are further divided into type II-A, II-B and II-C systems, where *cas1*, *cas*2, *cas9* and *trans*-activating crRNA (tracrRNA) are present in all the subtypes (Chylinski et al., 2013, 2014; Fonfara et al., 2014; Makarova et al., 2011). tracrRNA is a small RNA that contains an anti-repeat sequence complementary to the repeat sequences of pre-crRNAs (Deltcheva et al., 2011).

Bioinformatic studies demonstrated that two *csn2* variants, *i.e.* a shorter *csn2* (*csn2a*) that is always accompanied by the longer *cas9*, and a longer *csn2* (*csn2b*) that always coexists with the shorter *cas9* (Chylinski et al., 2014). Csn2 is the signature protein for the type II-A CRISPR-Cas systems (Makarova et al., 2011), and it is essential for spacer acquisition (Barrangou et al., 2007; Garneau et al., 2010; Heler et al., 2015; Wei et al., 2015a). Biochemical studies revealed that Csn2 exhibits a tetrameric ring-like structure (Ellinger et al., 2012; Koo et al., 2012; Lee et al., 2012; Nam et al., 2011) that could bind to the linear ends of dsDNA (Arslan

et al., 2013) and slide along the DNA in an energy-independent manner (Arslan et al., 2013). The structure and properties of Csn2 share similarity to the Ku protein from the non-homologous end joining (NHEJ) DNA repair pathway (Feldmann et al., 2000).

In addition to *cas1, cas2* and *cas9*, the type II-B systems encode *cas4* (Makarova et al., 2011, 2015). Cas4 has been shown to be necessary for spacer acquisition in several type I systems, as well as in *Sulfolobus solfataricus* and *P. furiosus* that encode several type I and type III systems together with different repeat families and an adaptation cassette (Kieper et al., 2018; Lee et al., 2018; Li et al., 2014; Liu et al., 2017; Rollie et al., 2018; Shiimori et al., 2017, 2018). Among the subtypes of the type II systems, the type II-C systems are minimal, as they only encode *cas1*, *cas2* and *cas9* (Chylinski et al., 2013, 2014; Fonfara et al., 2014).

1.3.2 Type II-A crRNA biogenesis

In the type II-A system, tracrRNA base-pairs with each repeat sequence of the pre-crRNA to form a tracrRNA anti-repeat:pre-crRNA repeat duplex, which will be bound to and stabilized by Cas9 (Deltcheva et al., 2011). RNase III recognizes and cleaves the tracrRNA:pre-crRNA duplexes within the anti-repeat sequences, resulting in intermediate tracrRNA:crRNA duplexes, of which the intermediate crRNAs consists of the 5′-repeat-spacer-repeat-3′ sequences (Deltcheva et al., 2011). The intermediate tracrRNA:crRNA duplexes are further processed by unknown RNase(s) into mature crRNAs, which are composed of 5′-spacer-repeat-3′ sequences (Deltcheva et al., 2011). Following the processing, Cas9 remains bound to the tracrRNA:crRNA duplex and forms the effector complex for interference (Deltcheva et al., 2011; Gasiunas et al., 2012; Jinek et al., 2012).

1.3.3 Type II-A CRISPR interference

Structural studies disclosed that Cas9 possesses two lobes, which are the α-helical recognition (REC) lobe and the nuclease (NUC) lobe. These two lobes are connected together by a flexible linker and a highly conserved arginine-rich bridge helix that interacts with the guide RNA. The NUC lobe is comprised of a HNH nuclease domain, a RuvC nuclease domain and a PAM-interacting domain at the C-terminus (Anders et al., 2014; Hirano et al., 2016; Jiang et al., 2015, 2016; Jinek et al., 2014; Nishimasu et al., 2014, 2015; Yamada et al., 2017). The PAM-interacting region of Cas9 is disordered in apo-Cas9, therefore apo-Cas9 is unable to

specifically bind DNA (Jinek et al., 2014). Structural studies revealed that the binding of the engineered dual-tracrRNA:crRNA, known as single-guide RNA (sgRNA) chimera, to Cas9 leads to conformational changes in the REC lobe and PAM-recognition region of Cas9, which transforms Cas9 from the inactive state (apo-Cas9) to the active state (sgRNA-bound Cas9) that can recognize the target DNA (Jiang et al., 2015). The critical element for this conformation activation is the seed sequence of the crRNA, which is composed of 10- to 12-PAM-adjacent nucleotides (nts) (Jiang et al., 2015; Jinek et al., 2012).

During interference, the activated guide RNA-bound Cas9 scans the target DNA for recognition of the PAM on the non-target strand (Sternberg et al., 2014). Upon PAM recognition, crRNA probes for sequence complementarity with the target DNA. The base-pairing of the crRNA with the target DNA strand will displace the non-target strand through local DNA unwinding and lead to the formation of an R-loop (Anders et al., 2014; Jinek et al., 2012; Mekler et al., 2017; Sternberg et al., 2014; Szczelkun et al., 2014). The stable R--loop formation and subsequently interference are more tolerated to the mismatches outside of the seed sequence (Cong et al., 2013; Jiang et al., 2013; Jinek et al., 2012; Sternberg et al., 2014). Upon R-loop formation, the HNH and RuvC nuclease domains of Cas9 cleave the target and the non-target DNA strands, respectively, which ultimately leads to a blunt double-stranded break (DSB) (Garneau et al., 2010; Jinek et al., 2012).

1.4 CRISPR adaptation (spacer acquisition)

CRISPR adaptation or spacer acquisition, the first stage of the CRISPR-Cas immunity, is critical for the generation of heritable immunological memory. Several steps are involved in spacer acquisition, which include MGE recognition, prespacer selection, prespacer processing and spacer integration. The core proteins of spacer acquisition are the almost ubiquitous Cas1 and Cas2 proteins (Koonin et al., 2017; Makarova et al., 2011, 2013, 2015; Shmakov et al., 2015), which were shown to form an integrase complex consisting of one Cas2 dimer spanning two Cas1 dimers in the type I-E and type II-A systems (Nunez et al., 2014; Xiao et al., 2017)3/11/19 7:38:00 PM. In this complex, Cas1 plays a catalytic role, whereas Cas2 plays a structural role in maintaining the Cas1-Cas2 complex (Nunez et al., 2014; Xiao et al., 2017). In the type I-E system of *E. coli*, Cas1 and Cas2 are the sole players in naïve spacer acquisition (Yosef et al., 2012), which refers to spacer incorporation when a pre-existing spacer against the

target is absent (Datsenko et al., 2012; Yosef et al., 2012). On the other hand, naïve spacer acquisition additionally requires Csn2, Cas9 and tracrRNA in type II-A systems of *S. pyogenes* and *S. thermophilus* (Heler et al., 2015; Wei et al., 2015a).

Thus far, our knowledge of spacer integration arises mainly from type I-E and type II-A systems. In view of this, the spacer acquisition mechanism of type I-E will be described alongside the one of type II-A, which is the focus of this thesis. The subtype-specific accessory proteins and variations of spacer acquisition mechanisms will also be addressed.

1.4.1 The provenance of prespacers

Spacer acquisition commences with the recognition of MGEs. While acquiring foreign DNA as a genetic memory for future defense might be life-saving, acquiring self-DNA could be life-threatening. Therefore, as described above, a preference of acquiring non-self DNA against self-DNA is crucial to prevent auto-immunity and ensure that the immunity is specifically targeting MGEs. The self- versus non-self discrimination by the homologous recombination machinery AddAB (for Gram-positive bacteria) and the RecBCD complex (for Gram-negative bacteria) has been described in the type II-A system of *S. pyogenes* (Modell et al., 2017) and the type I-E system of *E. coli* (Levy et al., 2015), respectively.

During DNA replication, DSBs often arise at the stalled replication forks. The RecBCD/AddAB complex is recruited to the DSBs, where the complex unwinds and degrades the dsDNA until a crossover hotspot instigator (Chi) site, which are short DNA motifs marking the DNA loci hotspot for the initiation of homologous recombination. The homologous recombination is commenced upon the loading of the RecA to the 3´-end overhang of the damaged DNA near the Chi site by RecBCD/AddAB. (Dillingham and Kowalczykowski, 2008; Smith, 2012; Wigley, 2013)

A plasmid-based spacer acquisition in the type I-E system of *E. coli* revealed that the sequences between the stalled replication forks and the Chi sites are the prespacer sampling hotspots (Levy et al., 2015). The authors showed that spacer acquisition was decreased in the individual deletion mutants of *recB*, *recC* and *recD*, and the prespacer sampling hotspots between the stalled replication fork and the Chi site were impaired in these mutants. Thus, it was proposed that the degradation fragments generated by the RecBCD could be employed by the Cas1-Cas2

complex for spacer integration (Levy et al., 2015). It was suggested that the degraded single-stranded DNA (ssDNA) products occasionally re-anneal and form dual-forked DNAs (dsDNA with 5′- and 3′-overhangs at both ends), which are the preferred prespacer structure for Cas1-Cas2 (Levy et al., 2015; Wang et al., 2015). Since there are fewer Chi sites on a plasmid compared to the *E. coli* chromosomes, a longer fraction of plasmid DNAs will be degraded by RecBCD until a Chi site is encountered (Levy et al., 2015). Therefore, this leads to a larger supply of the plasmid-derived prespacer substrates and favors spacer acquisition from the plasmid DNA over bacterial chromosomes (Levy et al., 2015). In addition, high-copy number plasmids undergo frequent replications, therefore there are more stalled replication forks on the plasmids, which results in a higher number of DSBs on these DNAs. A higher number of DSBs leads to a higher frequency of plasmid DNA degradation by RecBCD, therefore high-copy number plasmids are favored for spacer acquisition instead of bacterial chromosomes (Levy et al., 2015).

A phage infection study of the CRISPR-Cas type II-A system of *S. pyogenes* in a heterologous host *S. aureus*, demonstrated that the prespacers sampling hotspots are the DNA sequences between the cohesive end sites (cos sites) and the Chi sites (Modell et al., 2017). The cos site is the cohesive DNA ends of the linear viral DNA generated during the viral DNA packaging, and it allows the recircularization of the linear DNA upon DNA injection into the cell (Catalano et al., 1995). In the phage study of the type II-A system, the cos site manifests the free DNA end of the phage that initially enters the cell (Modell et al., 2017). An injected phage DNA could be degraded by the helicase-nuclease AddAB/RecBCD until a Chi site is encountered (Bobay et al., 2013; Dillingham and Kowalczykowski, 2008). In agreement with the observation that AddAB is required for efficient spacer acquisition in the type II-A system, it was proposed that AddAB supplies the DNA degradation products to the Cas1-Cas2 complex as prespacer substrates (**Figure 2A**) (Modell et al., 2017), which is also suggested for RecBCD in *E. coli* (Levy et al., 2015). Nevertheless, a very recent publication demonstrated that the helicase activity of the RecBCD is essential for promoting spacer acquisition in the type I-E system of *E. coli*, whereas the nuclease activity is not needed for spacer acquisition (Radovčić et al., 2018). Based on these findings, the authors suggested that the helicase-translocase activities of RecBCD promote spacer acquisition by removing the nucleoprotein complexes from the DNA damage sites, thereby allowing the access of Cas1-Cas2 for extracting the prespacer substrates. More studies are needed to clarify the roles of the homologous recombination machineries in spacer acquisition.

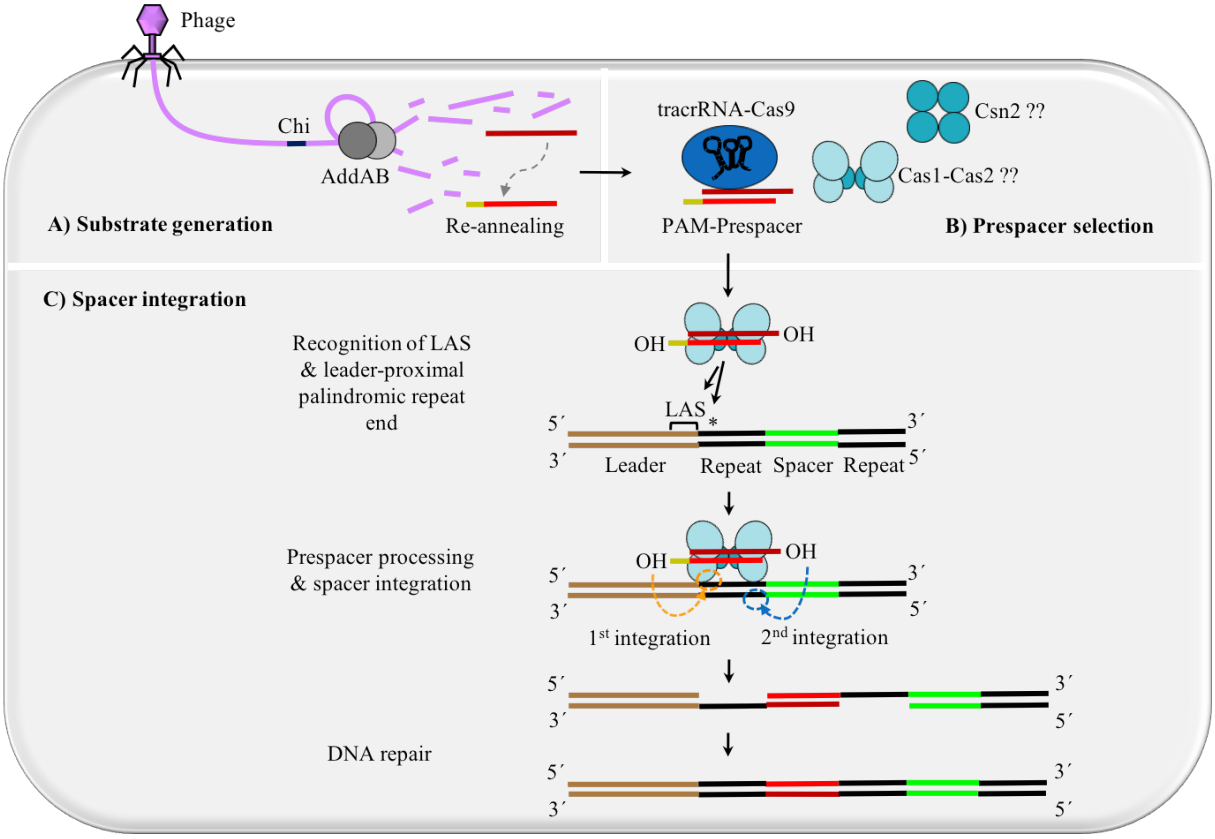


Figure 2. Spacer acquisition of type II-A systems.

(A) Upon phage challenge in Gram-positive bacteria, if the injected linear phage DNA is uncapped (or unprotected) by phage proteins, the phage DNA can be bound and degraded by the DNA repair complex AddAB until a Chi site is encountered. Some of the degraded ssDNA fragments may re-anneal and form dsDNA fragments with 3′-overhangs at both ends, which can serve as prespacer substrates for the acquisition machinery. **(B)** In the type II-A system, tracrRNA-bound Cas9 (tracrRNA-Cas9) selects the prespacers with canonical PAMs for spacer integration by Cas1-Cas2 complex. It remains unclear how the prespacer is transferred to the Cas1-Cas2 complex after the PAM recognition by tracrRNA-Cas9. The role of Csn2 in spacer acquisition is also not understood at this point. **(C)** Cas1-Cas2-prespacer complex recognizes the leader anchoring site (LAS) and leader-proximal end of the first repeat (*), and this enables preferential spacer integration at the leader-proximal CRISPR array. The 3′-OH of one strand of the prespacer carries out a nucleophilic attack on the leader-end of the first repeat (dotted arrow and circle in orange) and ligates the leader-end of the first repeat – an event known as the first integration. Subsequently, the 3′-OH of the other strand carries out a second nucleophilic attack and ligates on the spacer-end of the first repeat (dotted arrow and circle in blue), thereby resulting in the second integration event, *i.e.* full-site integration. The gap of the CRISPR array is presumably repaired by the DNA polymerase and ligase, and eventually results in the duplication of the first repeat.

1.4.2 The selection and processing of prespacers

Apart from the AddAB/RecBCD-mediated role in spacer acquisition, prespacer selection during spacer acquisition is often not a random process. In type I and type II systems, a prespacer with a PAM is preferentially selected by the adaptation machinery to ensure the acquisition of a functional spacer that could result in interference (Datsenko et al., 2012; Deveau et al., 2008; Mojica et al., 2009; Swarts et al., 2012).

1.4.2.1 The selection and processing of prespacers in the type II-A system

In the type II-A system, Cas9 is responsible for selecting a prespacer with a canonical PAM (**Figure 2B**) (Heler et al., 2015). This is supported by the observation that mutations on the PAM-recognition domain of *S. pyogenes* Cas9 led to random selection of prespacers that are not located next to a PAM. Furthermore, when the *S. pyogenes* Cas9 and *S. thermophilus* Cas9 from CRISPR3 locus (an ortholog of *S. pyogenes* Cas9) (Fonfara et al., 2014) were exchanged, the newly acquired spacers matched the PAMs that corresponded to the respective Cas9 ortholog, *i.e.* NGG PAM for *S. pyogenes* Cas9 (Jinek et al., 2012) and NGGNG PAM for *S. thermophilus* Cas9 from the CRISPR3 locus (Horvath et al., 2008).

It was shown that tracrRNA-bound Cas9 (tracrRNA-Cas9) is needed for crRNA biogenesis and interference (Deltcheva et al., 2011; Jinek et al., 2012). Likewise, the requirement of tracrRNA for spacer acquisition suggests a role for tracrRNA-Cas9 in spacer acquisition (Heler et al., 2015). However, tracrRNA-Cas9 does not appear to be involved in prespacer processing, as spacer acquisition could still be detected in catalytically inactive Cas9 (dead Cas9 or dCas9) (Heler et al., 2015; Wei et al., 2015a). The responsible player and the detailed mechanism for the prespacer processing still remains to be elucidated.

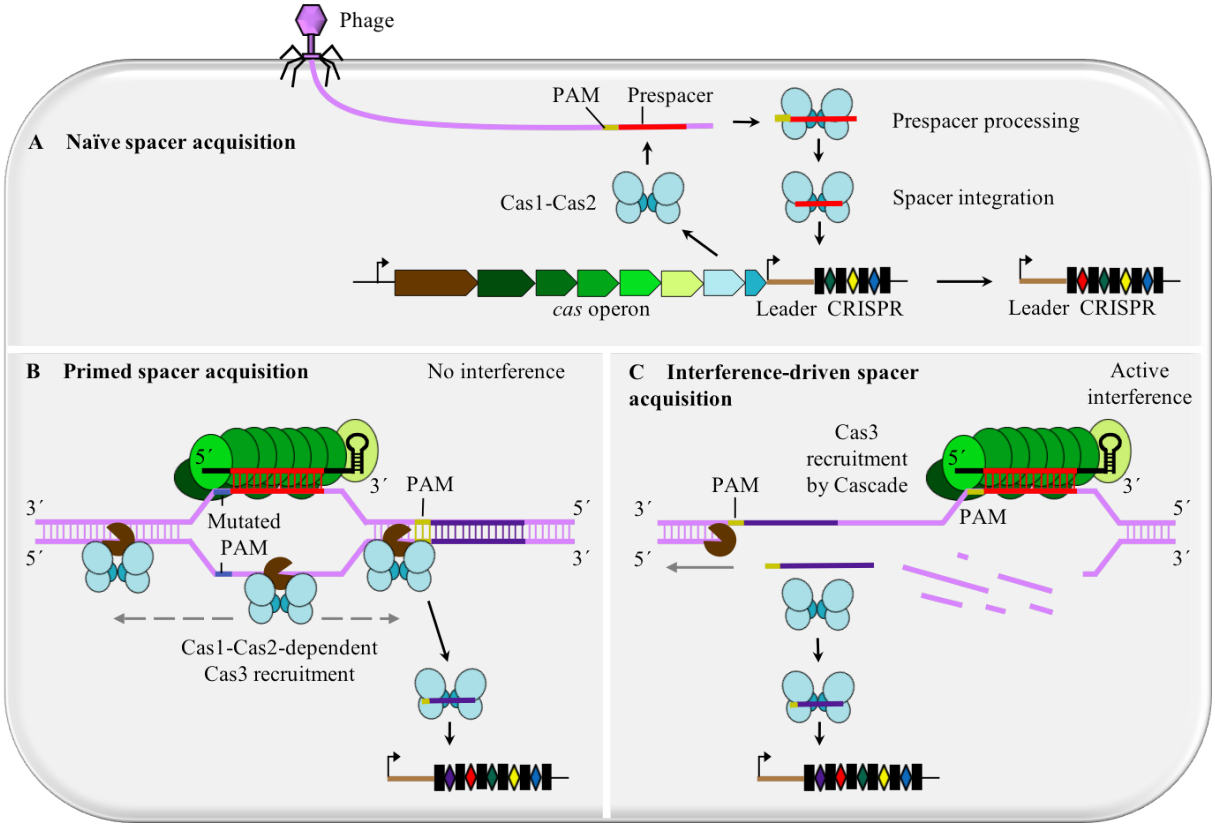


Figure 3. Spacer acquisition of type I systems.

(A) Naïve spacer acquisition occurs when there is no pre-existing spacer matching the invader. In type I-E systems, it was suggested that the degraded fragments of the foreign DNA generated by RecBCD complex (not illustrated here) can be used as prespacer substrates by the Cas1-Cas2 complex if the double-stranded substrates contain PAM. The prespacer is processed by Cas1-Cas2 prior to spacer integration into the leader-end of the CRISPR array. **(B)** Primed spacer acquisition occurs in the presence of a pre-existing spacer matching the target DNA. Mutation on the PAM or seed sequence of the protospacer abrogate interference, however, Cascade (green) can still bind to the target sequence with mutated PAM or seed sequence, but Cascade is unable to recruit the Cas3 nuclease (brown). In this case, the recruitment of Cas3 can be restored in the presence of Cas1-Cas2 complex. This non-canonical recruitment allows bidirectional translocation (grey dotted line with double arrows) of Cas3 together with Cas1-Cas2 complex without DNA cleavage. Upon identification of another canonical PAM, the sequence proximal to PAM can be used by Cas1-Cas2 as a prespacer substrate. **(C)** During interference-driven spacer acquisition, the DNA cleavage products generated by Cas3 nuclease could be captured by Cas1-Cas2 for spacer integration, thereby enlarging the spacer repertoire against a specific MGE. Unlike Cas1-Cas2-dependent recruitment of Cas3, Cascade recruits Cas3 upon target DNA identification. Cas3 cleaves the non-target DNA strand while translocating in 3′to 5′direction (grey arrow). Rectangles, repeats; diamonds, spacers. The colors of the spacers indicate variable MGEs-derived sequences. (Figure adapted from (Hille et al., 2018))

1.4.2.2 The selection and processing of prespacers in the type I-E system of *E. coli*

Interestingly, *in vivo* studies of the type I-E system of *E coli* demonstrated that, unlike to the type II-A system, the presence of Cas1 and Cas2 are sufficient for PAM recognition (**Figure 3A**) (Datsenko et al., 2012; Swarts et al., 2012). In line with this finding, a structural study also demonstrated that the Cas1 subunits of the type I-E Cas1-Cas2 complex preferably bind to the PAM-complementary sequence at the 3′-overhang (Wang et al., 2015). In addition to PAM, dual-forked DNA substrates with a minimum of 7-nt overhangs at both of the 3′-ends are also preferably bound by Cas1-Cas2 (Wang et al., 2015). When Cas1-Cas2 binds the dual-forked DNA, Cas1 twists the 3′-overhangs from the duplex DNA and stabilizes the dual-forked DNA substrate via two of its tyrosine residues (Wang et al., 2015). This structural arrangement allows the 3′-overhangs of the substrate to be located proximal to the active sites of Cas1 for processing, which leads to a 33-nt cleavage product that contains a 23-nt duplex DNA and 5-nt 3′-overhangs with 3′-OH at both ends. During the processing, parts of the PAM-complementary sequence are trimmed, thereby avoiding the integration of spacer with a PAM-sequence that could result in self-cleavage. In the Cas1-Cas2 complex, the distance of the two tyrosine residues of each of the Cas1 dimers allows the accommodation of a duplex DNA with the length of 22-23-bp. Hence, the assembly of the Cas1-Cas2 complex serves as a molecular ruler that pre-determines the length of the prespacers (Nunez et al., 2015a; Wang et al., 2015).

1.4.2.3 The selection and processing of prespacers in the type I-E system of *S. thermophilus* DGCC7710

Unlike the type I-E system of *E. coli*, in the type I-E system of *S. thermophilus* DGCC7710, Cas2 is fused to a DnaQ exonuclease domain, which allows to trim the 3′-overhangs of the prespacer with its 3′–5′ exonuclease activity to promote optimal spacer integration by the Cas1—Cas2-DnaQ complex (Drabavicius et al., 2018). In this in vitro study, efficient integration was observed when the 3′-ends possesses a pyrimidine deoxynucleotide.

1.4.2.4 The selection and processing of prespacers in the type I systems that encode Cas4

In several type I systems that encode Cas4, Cas4 has been reported to have a role in PAM recognitions and prespacer processing (Kieper et al., 2018; Lee et al., 2018; Rollie et al., 2018; Shiimori et al., 2018). Cas4 is a widespread protein that could be identified in the type I (I-A, I-B, I-C, I-D and I-U), II (II-B) and V (V-A, V-B and V-E) systems (Hudaiberdiev et al., 2017; Koonin et al., 2017; Makarova et al., 2015; Shmakov et al., 2015).

Cas4 possesses four conserved cysteine residues that allow its binding to the iron-sulfur clusters, and RecB-like nuclease motifs that are important for nuclease activity (Lemak et al., 2013, 2014; Zhang et al., 2012). DNA helicase, metal-dependent endonuclease and exonuclease activities have been reported for Cas4 (Lemak et al., 2013, 2014; Zhang et al., 2012).

In an *in vitro* experiment, Cas4 the type I-A system of *S. solfataricus*, was shown to be involved in the processing of the 3´-overhang of prespacer substrates in a PAM-dependent manner (Rollie et al., 2018). In the presence of Cas1-Cas2, Cas4 of the type I-C system of *Bacillus halodurans* promotes the processing of the 3´-overhang of prespacer in a PAM-dependent manner to ensure that only interference-favored prespacers will be integrated into the CRISPR array (Lee et al., 2018). In this regard, the nuclease activity of Cas1 also contributes to prespacer processing, although the nuclease activity of Cas4 would be sufficient (Lee et al., 2018). In agreement with that, an *in vivo* study in type I-D system (*Synechocystis sp.* 6803) also demonstrated that the nuclease activity of Cas4 is important for the processing of prespacers and promoting selection of spacers with a canonical PAM before spacer integration (Kieper et al., 2018).

A recent study showed that *Pyrococcus furiosus* encodes Cas4-1 and Cas4-2, which recognize a PAM and a NW motif, respectively, upstream and downstream of the prespacer; thereby it allows the nuclease activities of Cas4 to define the spacer length while processing the prespacer (Shiimori et al., 2018). *P. furiosus* encodes a type I-A Csa, a type I-G Cst and a type III-B Cmr effector complexes and seven CRISPR arrays with highly conserved leaders and repeats (Shiimori et al., 2017). In *P. furiosus*, *cas1*, *cas2* and *cas4-1* share the same locus with *cst*, *cmr* and *cas6* genes, whereas *cas4-2* is distant from other *cas* genes (Shiimori et al., 2017). It is still remains to be investigated whether Cas4 from the type II-B, V-A, V-B and V-E systems, are also involved in the prespacer processing.

1.4.3 Spacer integration into the CRISPR array

1.4.3.1 Recognition of the CRISPR array

The leader sequence (located at the upstream of the CRISPR array) polarizes the spacer integration predominantly at the leader-proximal CRISPR array, thereby maintaining the chronological order of the MGE invasions (Barrangou et al., 2007; Pourcel et al., 2005). In type II-A systems, the preference of spacer integration at the leader-proximal CRISPR array is determined by a short DNA motif named leader-anchoring site (LAS), which comprises 5 bp of the leader sequence immediately upstream of the CRISPR array (**Figure 2C**) (McGinn and Marraffini, 2016). In the type II-A system of *E. faecalis*, a LAS with a size of 4 bp is sufficient for determining the spacer integration at the leader-proximal CRISPR array (Xiao et al., 2017). Cas1-Cas2 uses the LAS as a landmark and inserts new spacers into the CRISPR array immediately downstream of the LAS (McGinn and Marraffini, 2016). However, when the LAS was mutated, inefficient ectopic spacer integration occured in the middle of the CRISPR array. Hence, the tolerance of Cas1-Cas2 to the mutations of the LAS allows flexible spacer integration at the non-leader-proximal sites of the CRISPR array, although the phage resistance is reduced possibly due to the lower expression of the crRNAs further from the leader (McGinn and Marraffini, 2016). In addition to the LAS recognition in the type II-A system, the *S. pyogenes* type II-A Cas1-Cas2 complex also recognizes the palindromic ends of the repeats via the Cas1 active sites (Wright and Doudna, 2016). Although the first integration could occur at either ends of the first repeat (leader-proximal repeat), the preference of spacer integration at the leader-end of the first repeat is directed by the concurrent recognition of both, the LAS and the palindromic end of the leader-end of the first repeat by Cas1-Cas2 (**Figure 2C**) (McGinn and Marraffini, 2016; Wright and Doudna, 2016; Xiao et al., 2017). A recent study demonstrated that the middle part of the repeat in *E. faecalis* is also important for spacer acquisition, as spacer acquisition was abrogated when the middle part of the repeat sequence of *E. faecalis* was replaced with the middle part of the repeat sequence of *S. pyogenes* (Xiao et al., 2017).

Unlike the type II-A system, the recognition of the leader-proximal CRISPR array in the type I-E system of *E. coli* relies on the non-CRISPR protein named integration host factor (IHF) (Nunez et al., 2015a, 2016; Yoganand et al., 2017). Upon binding of IHF on the leader

sequence, the leader sequence bends into a U-shaped DNA structure that facilitates the recognition of the leader-repeat boundary by Cas1-Cas2, which positions the prespacer pre-loaded on the Cas1-Cas2 close to the leader-repeat boundary for integration (Wright et al., 2017).

1.4.3.2 Spacer integration

In the type II-A systems, upon binding of Cas1-Cas2 to the LAS-repeat sequence, the palindromic-repeat recognition loop of Cas1 undergoes a conformational change, which allows the 3´-OH ends of the prespacer to be positioned close to the catalytic sites of Cas1 before the first half-site integration (the integration of one strand of the prespacer at one end of the repeat) (Xiao et al., 2017). During the first half-site integration in both type II-A (**Figure 2C**) and type I-E systems, Cas1 catalyzes the nucleophilic attack of the 3´-OH end of the prespacer on the leader-end of the repeat, thereby attaching the 3´-overhang of the prespacer to the leader-end of the repeat (Nunez et al., 2015a; Wright and Doudna, 2016; Xiao et al., 2017).

In both type II-A and type I-E systems, the second integration relies on the recognition and binding of Cas1-Cas2 at the palindromic repeat motif at the spacer-end of the repeat, which subsequently leads to the bending of the repeat sequence, thereby allowing the accessibility of Cas1 to the second integration site at the spacer-end of the repeat (Goren et al., 2016; Wright et al., 2017; Xiao et al., 2017). In addition, a prespacer with the proper length is also important for the second integration (Wright and Doudna, 2016). During the second nucleophilic attack, the 3´-OH of another strand of the prespacer is ligated to the spacer-end of the repeat, hence, full spacer is completely inserted into the CRISPR array (**Figure 2C**) (Nunez et al., 2015a, 2016; Wright and Doudna, 2016). If the mentioned criteria are not met, the full-site integration could be restrained, probably through the reversal of the first half-side integration reaction by the acquisition machinery, or the elimination of the half-side integration intermediate product by DNA repair proteins (Wright and Doudna, 2016). Subsequently, the gaps of the CRISPR array are presumed to be repaired by DNA polymerase and ligase, which leads to spacer integration and repeat duplication.

The proper orientation of the integrated spacer in the CRISPR array ensures the production of crRNA that can bind to the protospacer with a PAM and thus to enable interference. In the type I-E system of *E. coli*, the proper orientation of the newly integrated spacer is determined by the presence of the partial PAM on the prespacer, which is trimmed possibly after the binding

of Cas1-Cas2 to the leader-repeat boundary and prior to spacer integration (Shipman et al., 2016; Shmakov et al., 2014; Wang et al., 2015). In *P. furiosus*, Cas4-1 and Cas4-2 maintain the orientation of the integrated spacer: (1) via processing the 5′-NGG PAM (by Cas4-1) and the 3′-NW motif (by Cas4-2) of the prespacer prior to the integration, and (2) by staying associated with the spacer during the integration process (Shiimori et al., 2018).

The *in vitro* spacer integration in the type II-A system of *S. pyogenes* system does not require Csn2 (Wright and Doudna, 2016), which was shown to be essential for spacer acquisition *in vivo* (Barrangou et al., 2007; Garneau et al., 2010; Heler et al., 2015; Wei et al., 2015a). Csn2 was shown to co-purify with Cas1, Cas2 and Cas9, which are also critical for spacer acquisition (Heler et al., 2015) (Heler 2015). Moreover, size-exclusion chromatography (SEC) showed the direct interaction of Csn2 with Cas1 and Cas9, respectively (Ka et al., 2016, 2018). Based on the observation that the N-terminal interacting site of Cas1 with Csn2 is located close to the DNA end of the prespacer, Csn2 was suggested for its role associated with prespacer generation or a role as a scaffold protein for connecting other Cas proteins (Ka et al., 2018; Wright and Doudna, 2016) or host factors. However, the roles of Csn2 in spacer acquisition still remains to be clarified experimentally.

1.4.4 Primed spacer acquisition

Mutating the protospacers or PAMs is one of the strategies applied by MGEs to escape CRISPR-Cas immunity (Deveau et al., 2008; Fineran et al., 2014; Semenova et al., 2011). From the perspective of CRISPR-Cas immunity, harboring multiple-spacers targeting a single-MGE could counter-combat mutated MGEs (Paez-Espino et al., 2013; Xue et al., 2015) and hamper the proliferation of MGE escape mutants effectively (Andersson and Banfield, 2008; van Houte et al., 2016). With this regard, a pre-existing spacer (in the CRISPR array) that corresponds either to a mutated protospacer or a mutated PAM, can accelerated the spacer uptake through a process named primed spacer acquisition (also known as priming) (**Figure 3B**) (Datsenko et al., 2012; Fineran et al., 2014; Li et al., 2014; Richter et al., 2014; Swarts et al., 2012). In contrast to the primed spacer acquisition, naïve spacer acquisition refers to the spacer uptake from newly confronted MGEs (**Figure 3A**), therefore there is no pre-existing spacer against the newly confronted MGEs, and it shows relatively lower rate of spacer uptake in comparison to primed spacer acquisition (Yosef et al., 2012). To date, primed spacer acquisition has only been described in the type I CRISPR-Cas systems, and it remains to be studied whether other types

of CRISPR-Cas systems also adopt this counter-strategy against MGE mutants. Among the type I systems, primed spacer acquisition in the type I-E system of *E. coli* is the most extensively studied. Thereby, type I-E primed acquisition will be the focus here.

In the type I-E system, Cas1, Cas2 and the interference machinery, *i.e.* Cas3 and Cascade, are required for primed spacer acquisition (Datsenko et al., 2012; Li et al., 2014; Richter et al., 2014; Savitskaya et al., 2013; Swarts et al., 2012). When the target DNA matches the crRNA of the Cascade, Cascade binds to the protospacer and recruits Cas3 to cleave the MGE (Hochstrasser et al., 2014; Mulepati and Bailey, 2011; Sinkunas et al., 2011, 2013; Westra et al., 2012). When CRISPR interference is abrogated by the mutations in the PAM or seed sequence of the protospacer, Cascade binds to the target DNA but fails to recruit Cas3 (**Figure 3B**) (Blosser et al., 2015; Redding et al., 2015; Xue et al., 2016). If Cas1 and Cas2 are present, however, Cas3 can still bind to the target DNA and translocate bi-directionally along the DNA, presumably together with Cas1 and Cas2, searching for a prespacer with a canonical PAM for integration (Redding et al., 2015). In line with this, an adaptation complex constituted by Cas1 and Cas2-Cas3 fusion protein was reported in the type I-F system of *Pectobacterium atrosepticum* (Fagerlund et al., 2017; Richter et al., 2012) and *Pseudomonas aeruginosa* (Rollins et al., 2017), suggesting similarity between the primed acquisition mechanisms of the type I-F and the type I-E systems.

1.4.5 Interference-driven spacer acquisition

A variety of primed spacer acquisition known as interference-driven spacer acquisition also promotes rapid spacer acquisition in the presence of pre-existing spacers (Künne et al., 2016; Staals et al., 2016). However, as suggested by the name, interference-driven spacer acquisition is promoted by CRISPR interference (Künne et al., 2016; Staals et al., 2016), which is contrary to the mutation-stimulated primed spacer acquisition (Datsenko et al., 2012; Fineran et al., 2014; Li et al., 2014; Richter et al., 2014; Swarts et al., 2012). Studies in the type I-E system of *E. coli* and type I-F of *P. atrosepticum* demonstrated that the DNA cleavage products generated by Cas3 during interference, could be used by the Cas1-Cas2 for spacer acquisition, resulting in multiple spacers targeting the same MGEs (**Figure 3C**) (Künne et al., 2016; Staals et al., 2016). It is suggested that the ssDNA fragments generated by Cas3 re-annealed into various partial duplexes, including 3′- and/or 5′-overhangs, with an intermediate spacer length and 3′-PAM enrichments (Künne et al., 2016). Among the pool of prespacers with various

structures, Cas1-Cas2 shows binding preference to the partial duplexes with 3´-overhangs and 3´-PAMs and further trims the substrates before integration (Nunez et al., 2015b; Wang et al., 2015). Thus, cooperation between the adaptation and interference machineries allows Cas1-Cas2 to reuse the DNA degradation products for spacer integration, thereby, expanding the spacer repertoire. Interference-driven spacer acquisition promotes diverse spacer repertoires, which allows efficient interference against the same phage (Paez-Espino et al., 2013; Xue et al., 2015) and reduces the possibility of the propagation of escaped phages (van Houte et al., 2016; Künne et al., 2016; Staals et al., 2016). Both primed and interference-driven spacer acquisitions have only been reported in type I systems thus far, implying that different types of CRISPR-Cas systems adopt different modes of acquisition strategies that serve the best for their survival and fitness.

1.4.6 Reverse transcription spacer acquisition

While spacers are commonly acquired from foreign DNA across numerous types of CRISPR-Cas systems, both DNA and RNA could be acquired as spacers in the type III-B system of *Marinomonas mediterranea*. The type III-B system of *M. mediterranea* possesses a fusion protein of Cas1 with reverse transcriptase (RT-Cas1), which together with Cas2 inserts an RNA prespacer into the leader-proximal repeat, and subsequently reverse transcribes the inserted RNA into a cDNA spacer via the RT domain of Cas1. In addition, RT-Cas1-Cas2 can also incorporate a DNA prespacer into the CRISPR array via the conventional mode (Silas et al., 2016).

1.5 Co-evolution of phages and prokaryotes

The arm race between prokaryotes and phages is an endless process and the driving force for both parties to constantly evolve the defense and counter-defense strategies against each other. Similar to other prokaryotic defense systems, phages have also evolved a couple of CRISPR-Cas evasion strategies which were characterized. However, there are probably many more of these anti-CRISPR strategies awaiting to be discovered.

As a counter CRISPR-Cas strategy, some phages have evolved anti-CRISPR (*acr*) genes to directly hinder the CRISPR interference mechanism (Pawluk et al., 2017). The anti-CRISPR

proteins encoded in the temperate phages of *P. aeruginosa* demonstrated inhibitory activities against the type I-F system (Pawluk et al., 2017). For example, AcrF1 and AcrF2 interact with the Csy complex (crRNA-effector of type I-F system) subunits, thereby interfering with the binding of Cascade to the target DNA (Bondy-Denomy et al., 2015; Chowdhury et al., 2017; Peng et al., 2017), whereas AcrF3 prohibits the recruitment of the Cas3 effector for target DNA cleavage (Bondy-Denomy et al., 2015). In the type II-C system of *Neisseria meningitides*, AcrIIC prevents CRISPR interference by binding to Cas9 and blocking DNA cleavage *in vitro* (Pawluk et al, 2016, Cell). The type II-A anti-CRISPR protein, AcrIIA4, inhibits the PAM-interacting residues of Cas9 of *Listeria monocytogenes* (Dong et al., 2017; Rauch et al., 2017; Shin et al., 2017; Yang and Patel, 2017). A very recent study revealed that the AcrIIA5 and AcrIIA6 proteins inhibit Cas9 of the type II-A systems of *S. thermophilus* (CRISPR1 and CRISPR3) and *S. pyogenes*, however, the detailed mechanism remains to be investigated (Hynes et al., 2017, 2018).

While anti-CRISPR systems are protein-based strategies that have evolved in some phages over a longer timescale, some phages also use a random mutation strategy that depends on sequence specificity to evade CRISPR-Cas systems. With this regard, escaped mutants hamper CRISPR interference via the introduction of mutations in the PAM or seed sequence of the protoscaper (Deveau et al., 2008; Fineran et al., 2014; Semenova et al., 2011). However, the type III-A system of *S. epidermidis* is more tolerant towards the mutations in the protospacer and proximal regions due to its wide mismatches-tolerant, and PAM- and seed sequence-independent features (Pyenson et al., 2017). As a counter-strategy to escaped mutants, primed spacer acquisition has evolved in some type I systems, which boosts spacer acquisition when interference is interrupted by PAM or seed sequence mutations (Datsenko et al., 2012; Fineran et al., 2014; Li et al., 2014; Richter et al., 2014; Swarts et al., 2012). Primed spacer acquisition promotes spacer diversity in bacterial populations and subsequently relieves the evasion rate of the phage mutants (van Houte et al., 2016). On the other hand, *S. thermophilus* co-evolution studies revealed that multi-phage infection enhances the persistence duration of the phages (Paez-Espino et al., 2015). Furthermore, recombination of two phages were reported to diversify the phage genetic contents, consequently enhancing the chance to evade CRISPR-Cas systems (Paez-Espino et al., 2015).

Interestingly, instead of being sabotaged by the CRISPR-Cas systems, some phages exploit the CRISPR-Cas systems for their own benefits. For example, some of the *Campylobacter jejuni*

phages encode Cas4 (Siringan et al., 2014) that is absent in the type II-C system of *C. jejuni*, which encodes only *cas9*, *cas1*, *cas2* and tracrRNA (Chylinski et al., 2013, 2014). The phage-encoded Cas4 allows the uptake of spacers that are targeting solely *C. jejuni* genome (Siringan et al., 2014). Furthermore, the type I-F CRISPR-Cas system encoded by *Vibrio cholerae* phages is exploited to target other host phage defense systems on the genomic islands (Seed et al., 2013).

1.6 CRISPR-Cas systems and DNA repair pathways

Despite the fact that the CRISPR-Cas system is well known for its role in the adaptive defense system, CRISPR-Cas was initially proposed for functioning in DNA repair or chromosomal segregation pathways (Makarova et al., 2002). In line with the initial hypothesis, studies of the type I-E system of *E. coli* revealed the physical and genetic interaction between Cas1 and proteins from the DNA repair and homologous recombination pathways, including RecB, RecC and RuvB (Babu et al., 2011). Some of the DNA substrates of the Cas1 nuclease activity share similar structures with the intermediate products derived from the DNA repair and homologous recombination pathways, for example, Holiday junctions, replication forks and 5'-flaps (Babu et al., 2011; Nunez et al., 2015b; Wiedenheft et al., 2009). Furthermore, studies of the type I-E system of *E. coli* and type II-A system of *S. pyogenes* suggested that the Cas1-Cas2 complex possibly employs the degradation products generated by the RecBCD or AddAB complex for spacer acquisition (Levy et al., 2015; Modell et al., 2017). However, a very recent study reported that the nuclease activity of RecBCD is not required for spacer acquisition, instead, the helicase activity of RecBCD is essential for spacer acquisition (Radovčić et al., 2018).

In vivo, a *cas1* deletion mutant of *E. coli* is more sensitive to DNA damage caused by genotoxic agent mitomycin C (MMC) and ultraviolet irradiation. Correspondingly, Cas1 is also detected at the DNA damage sites in MMC-treated *E. coli*. Furthermore, a *cas1* deletion mutant demonstrates abnormal cells elongation after MMC treatment, implying a role of Cas1 in cell division and chromosomal segregation (Babu et al., 2011). In agreement with the studies in *E. coli*, MMC-induced DNA damage in *Acinetobacter baylyi* ADP1 that encodes a type I-F CRISPR-Cas system, increased the expression of *cas1*, as well as other *cas* genes, such as *cas6*, *csy2* and *csy3* (Hare et al., 2014). In type I-A of *Pyrococcus furiosus*, *cas* genes were upregulated upon exposure to gamma irradiation (Williams et al., 2007). Additionally, nalidixic

acid-induced SOS response causes accumulation of crRNA (Klaiman et al., 2014). Another study in *E. coli* revealed that the DNA repair proteins RecG, PriA and DNA polymerase I are required for primed adaptation, whereas RecB and DNA polymerase I are needed for naïve adaptation (Ivancic-Bace et al., 2015). A recent study showed that the non-homologous end-joining (NHEJ) system does not impact CRISPR immunity, and *csn2* of type II-A systems restrains the co-occurrence of the NHEJ system in bacteria (Bernheim et al., 2017). Despite these initial experimental evidences showing the connection between CRISPR-Cas systems and DNA repair pathways, the complex interactions between these systems and the detailed mechanisms involved are still yet to be elucidated.

1.7 The type II-A CRISPR-Cas systems of *Streptococcus pyogenes* SF370 and *Streptococcus thermophilus* LMD-9

1.7.1 Streptococcus pyogenes

1.7.1.1 General characteristics and pathogenesis of *S. pyogenes*

S. pyogenes, also known as group A *Streptococcus* (GAS), is a Gram-positive bacterium and a human pathogen that causes a wide spectrum of infections such as streptococcal toxic-shock syndrome, necrotizing fasciitis, pharyngitis and impetigo (Carapetis et al., 2016; Walker et al., 2014). While *S. pyogenes* is still susceptible to penicillin and cephalosporin treatments, there is a growing global concern of resistance to other antibiotics such as macrolides and clindamycin. Hitherto, an effective and safe vaccine to prevent *S. pyogenes* infections is still needed (Walker et al., 2014).

1.7.1.2 CRISPR-Cas loci in *S. pyogenes* SF370

S. pyogenes SF370 (also known as M1 GAS) possesses two CRISPR-Cas loci, *i.e.* the CRISPR01 locus, which belongs to the type II-A system; and the CRISPR02 locus, which belongs to the type I-C system (Deltcheva et al., 2011; Nozawa et al., 2011). The CRISPR locus

of the type II-A system of *S. pyogenes* SF370 showed expression, whereas the CRISPR locus of the type I-C system appears to be not expressed (Deltcheva et al., 2011). The type II-A system of *S. pyogenes* SF370 encodes tracrRNA, *cas9*, *cas1*, *cas2* and *csn2* and harbors a CRISPR array containing 7 repeats interspaced with 6 spacers (**Figure 4**). Spacer 6 does not match to any sequence, whereas spacers 1 to 5 share high sequence similarity to prophage sequences, *i.e.* endopeptidase, superantigen (*speM*), adenine-specific methyltransferase, hyaluronidase and phage protein (Deltcheva et al., 2011). The first spacer in the type II-A CRISPR array matches the prophage-encoded Spy_0700 gene existing in the genome of SF370 strain. However, the PAM corresponding to the first spacer is mutated (Deltcheva et al., 2011), presumably to avoid autoimmunity.

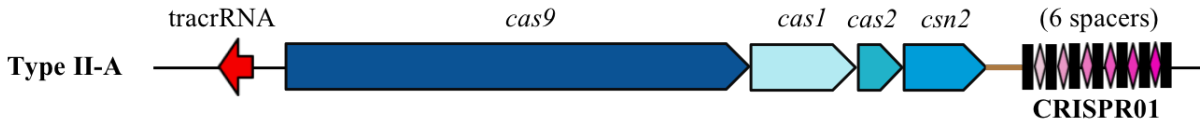


Figure 4. Type II-A CRISPR-Cas locus in *S. pyogenes* SF370.

Type II-A CRISPR-Cas systems encode four *cas* genes, *i.e. cas9*, *cas1*, *cas2*, *csn2*, and a tracrRNA. Brown line, leader; rectangles, repeats; diamonds, spacers.

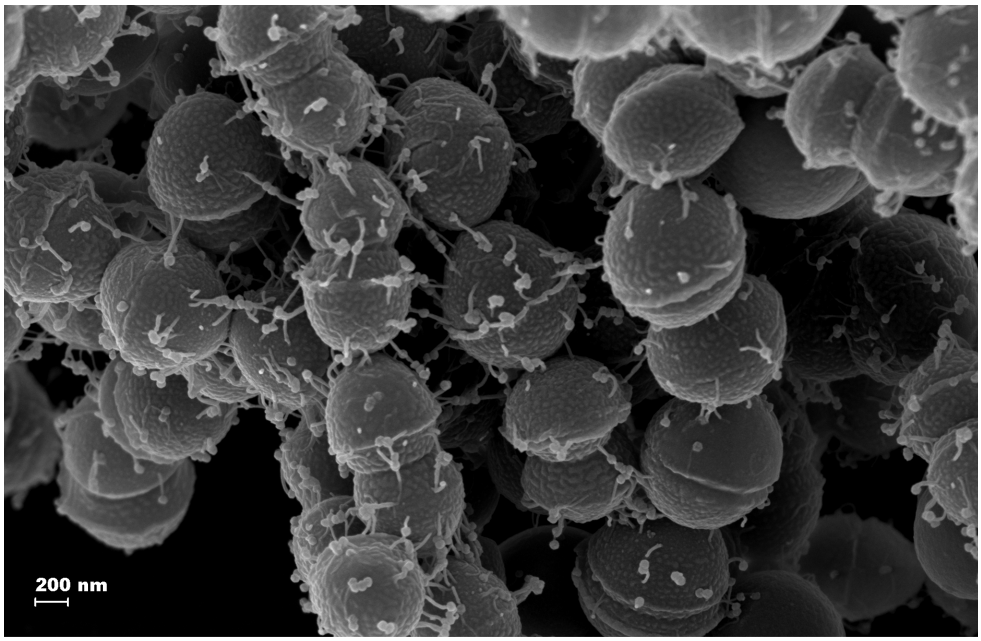
1.7.2 Streptococcus thermophilus

1.7.2.1 General characteristics and the industrial values of *S. thermophilus*

Streptococcus thermophilus is classified as lactic acid bacteria, which ferment carbohydrate to lactate for metabolic energy generation. Unlike the other pathogenic *Streptococci* including *S. pyogenes*, *S. thermophilus* has followed a different evolutionary path, and has lost many of the virulence-associated properties. Thus, it is regarded as a 'generally recognized as safe' species. On the other hand, *S. thermophilus* has gained industrially valuable features such as polysaccharide biosynthesis, bacteriocin biosynthesis, RM systems or oxidative stress endurance (Bolotin et al., 2004; Hols et al., 2005). *S. thermophilus* is regarded as the second most valuable bacteria in the dairy industry after *Lactococcus lactis* due to its extensive use as a starter culture in the production of yogurt and cheeses (Hols et al., 2005). Furthermore,

S. thermophilus, *Lactobacillus* sp., *Bifidobacterium* sp., and *Saccharomyces boulardii* are commonly used microorganisms in probiotics (Kligler and Cohrssen, 2008).

A



B

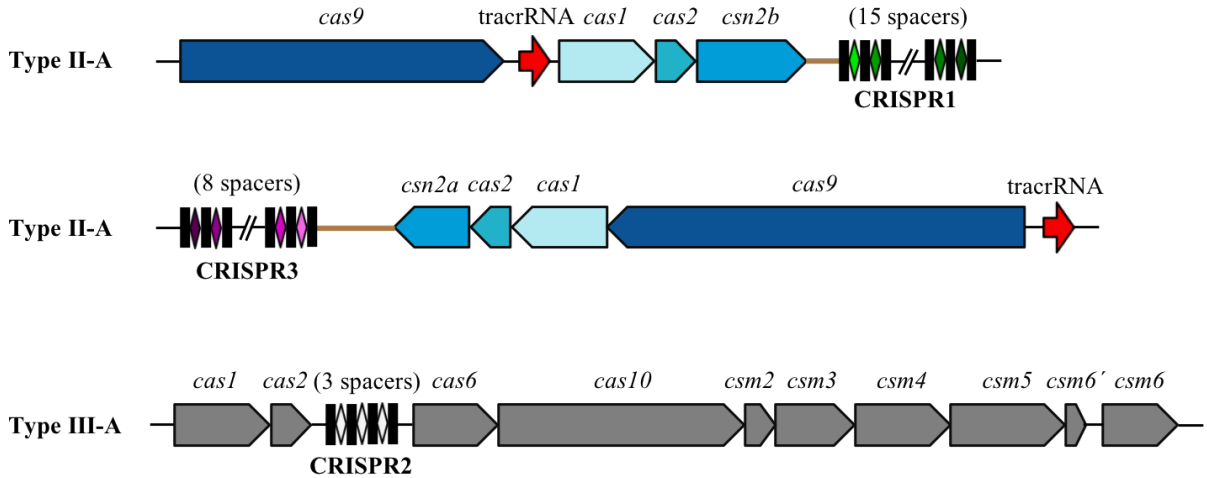


Figure 5. *S. thermophilus* LMD-9 and phage DT1.

(A) Scanning electron microscopy of *S. thermophilus* LMD-9 infected by phage DT1 at the multiplicity of infection (MOI) of 100 with the adsorption time of 1 minute at room temperature (Photograph courtesy of Manfred Rohde, Helmholtz Centre for Infection Research, Braunschweig). **(B)** CRISPR-Cas loci in *S. thermophilus* LMD-9. Brown line, leader; rectangles, repeats; diamonds, spacers.

1.7.2.2 Bacteriophage infection in *S. thermophilus*

During fermentation process in the dairy industry, phage infection in the starter cultures (**Figure 5A**) can cause huge problems in the dairy production and financial loss, as it could be an obstruction for the milk acidification by the starter cultures in the fermentation tanks (Brüssow, 2001; Quiberoni et al., 2010). Additionally, the endurance of phages to pasteurization and their airborne propagation leads to persistence of phages in the dairy production facilities. Hence, several bacterial defense systems have been exploited in dairy manufacturing industry to combat phage infection in the starter culture and minimize the negative impacts of phage infection during the manufacturing process (Barrangou et al., 2013). CRISPR-Cas systems can be harnessed to improve phage resistance of *S. thermophilus* along with other defense systems or in combination with numerous strain rotations to protect the dairy manufacturing process (Barrangou et al., 2013).

1.7.2.3 CRISPR-Cas loci in *S. thermophilus* LMD-9

S. thermophilus LMD-9 contains three CRISPR loci in its genome, namely the CRISPR1 and CRISPR3 loci that belong to type II-A systems, and the CRISPR2 locus that belongs to type III-A system (**Figure 5B**). Both type II-A CRISPR loci were shown to be active in spacer acquisition, whereas CRISPR2 of the type III-A system seems to be inactive (Horvath et al., 2008). The type III-A system of *S. thermophilus* LMD-9 contains the almost universal *cas1* and *cas2* genes, *cas6* that is important for crRNA biogenesis, and other *cas* genes encoding the subunits that form the Csm complex, namely *cas10*, *csm2*, *csm3*, *csm4*, *csm5* and *csm6* (Horvath et al., 2008). On the other hand, *cas9*, *cas1*, *cas2*, *csn2* and tracrRNA are encoded in the type II-A systems of the CRISPR1 and CRISPR3 loci (Horvath et al., 2008), of which the marked differences are CRISPR1 that contains short *cas9* and long *csn2* (*csn2b*), while CRISPR3 contains long *cas9* and short *csn2* (*csn2a*) (Chylinski et al., 2013). The CRISPR array of the CRISPR1 locus contains 15 spacers, whereas the CRISPR array of CRISPR3 locus contains 8 spacers (Horvath et al., 2008).

2 Aims of the Thesis

During CRISPR adaptation, a short piece of genetic material is acquired from the invading MGE and integrated into the CRISPR array as a heritable genetic memory to defend the prokaryotes against the re-invasion of the cognate MGEs. Spacer acquisition was first reported in the type II-A system of *S. thermophilus* in 2007 (Barrangou et al., 2007), however, the detailed molecular mechanism of spacer acquisition and the elements needed for the type II-A system were not understood at the time. In type I-E systems of *E. coli*, the almost universal Cas1 and Cas2 proteins are sufficient to activate naïve spacer acquisition *in vivo* (Yosef et al., 2012), and the mechanism was presumed to be similar in the type II-A system. When a *S. thermophilus* strain with deactivated *csn2* was challenged with lytic phage, spacer acquisition was abrogated, suggesting a role of Csn2 in this stage (Barrangou et al., 2007). The aim of my thesis was to characterize the molecular mechanism and the requirements for type II-A spacer acquisition. In addition to the implicated Cas1, Cas2 and Csn2, we were interested in investigating whether Cas9, tracrRNA and other non-Cas proteins were also involved in type II-A spacer acquisition. Furthermore, we planned to elucidate the role of every single element essential for spacer acquisition. This included examination of the direct protein-protein interactions among these proteins and with external co-factors, identification of the interacting regions and investigation of the significance of these interactions in spacer acquisition or other cellular processes. This thesis also aimed to examine whether there is any crosstalk in spacer acquisition between the Cas proteins from different CRISPR-Cas loci.

3 Results

3.1 Spacer acquisition in the type II-A CRISPR-Cas system of *S. pyogenes* SF370

At beginning of my PhD project, the understanding of the spacer acquisition mechanism in CRISPR-Cas immunity was limited. Cas1 and Cas2 was shown to be sufficient for spacer acquisition in the type I-E system of *E. coli* (Datsenko et al., 2012; Yosef et al., 2012). The essentiality of Cas1 and Cas2 in spacer acquisition were presumed to be same in other CRISPR-Cas types, including the type II-A systems, due to the high conservation of Cas1 and Cas2 in the CRISPR-Cas systems (Makarova et al., 2011), and their dispensability in crRNA biogenesis and interference (Barrangou et al., 2007; Bhaya et al., 2011; Brouns et al., 2008). Csn2 was shown to be additionally needed for spacer acquisition in the type II-A systems of *S. thermophilus* (Barrangou et al., 2007). We started our study by examining the significance of all CRISPR-associated elements, *i.e.* Cas1, Cas2, Csn2, Cas9 and tracrRNA, of the type II-A system in spacer acquisition. We then investigated whether any additional element was also required for the adaptation process.

3.1.1 The heterologous type II-A CRISPR-Cas system of *S. pyogenes* is established in *E. coli* BL21-AI

In order to genetically and biochemically characterize the spacer acquisition mechanism of *S. pyogenes* type II-A system, the type II-A system was cloned into plasmids and investigated in *E. coli* BL21-AI. This *E. coli* strain does not encode endogenous *cas* genes which makes it a suitable host for heterologous studies (Yosef et al., 2012). *E. coli* BL21-AI encodes T7 RNA polymerase that is regulated by L-arabinose-induced promoter, thereby allowing inducible expression of the genes under a T7-*lac* promoter of a plasmid.

The type II-A system was separated into two parts and cloned into two different plasmids. This aimed to separate the CRISPR array from *cas1*, *cas2* and *csn2* genes, to allow antibiotics selection of the clones that have lost the plasmids harboring the three *cas* genes, as result of

spacer acquisition. The combination of *cas1-cas2-csn2* genes was cloned into the pCDF-DUET vector, whereas tracrRNA-Leader-CRISPR was cloned into the pEC85. The expression of *cas1*, *cas2*, *csn2* was controlled by the inducible T7-*lac* promoter, whereas the expression of tracrRNA and CRISPR was regulated by their native promoters.

We confirmed the transcription of *csn2* in *E. coli* BL21-AI via semi-quantitative reverse transcription polymerase chain reaction (RT-PCR) (**Figure 6**). Since, *cas1*, *cas2* and *csn2* genes are located in the same operon and controlled by the same inducible promoter, the transcription of *csn2* is sufficient to show the transcription of the whole operon. The expression of the heterologous genes was also observed in the samples without induction, which might be due to leaky expression. The Cas proteins translation was not verified by Western Blot analysis due to the lack of antibodies against the Cas proteins at that time. After confirming the transcription of the *cas* operon with *cas1*, *cas2* and *csn2*, two different spacer acquisition setups were tested in the heterologous system, which were the plasmid-based acquisition assay and phage-based acquisition assay.

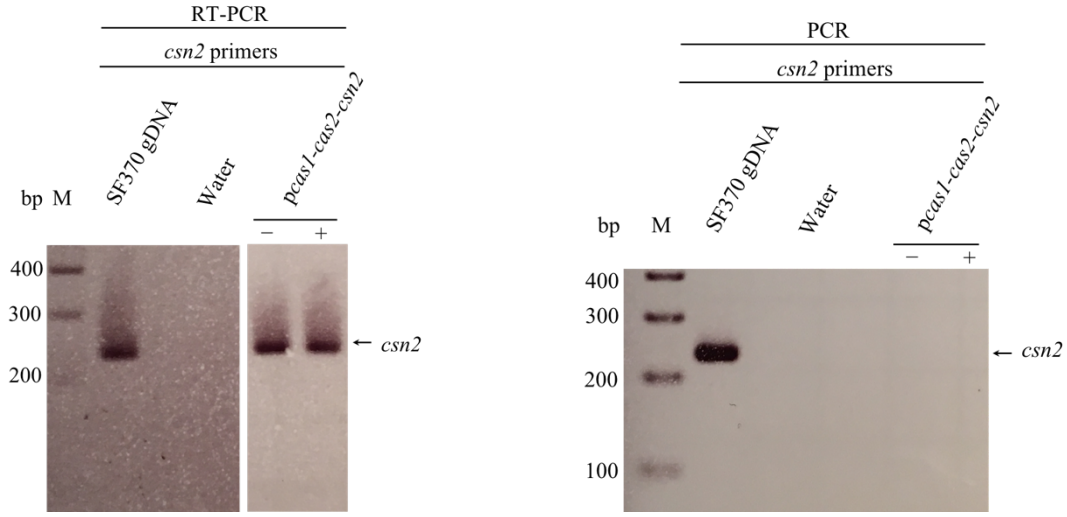


Figure 6. Validation of *csn2* gene expression in the heterologous type II-A CRISPR-Cas system of *S. pyogenes* SF370 in *E. coli* BL21-AI.

The agarose gel picture of semi-quantitative RT-PCR shows the expression level of *csn2* gene of pCDF-DUETΩ*cas1-cas2-csn*2. The expression of *cas* genes was induced with 1 mM IPTG and 0.2% arabinose for 2 hours. A PCR reaction was additionally performed to confirm that the RNA templates were free from DNA contamination. pCDF-DUET is abbreviated as "p" in the picture. SF370 gDNA, genomic DNA of *S. pyogenes* SF370; "–", without induction; "+", with induction; M, GeneRuler™ 100-bp DNA Ladder (Thermo Scientific).

3.1.1.1 Plasmid-based spacer acquisition in the heterologous system of *S. pyogenes*

We asked the question whether *cas1, cas2*, *csn2* and tracrRNA are sufficient for spacer acquisition in the type II-A system. To address this question, we induced the expression of *cas1, cas2* and *csn2* with isopropylthio-β-d-galactoside (IPTG) and arabinose in *E. coli* BL21-AI that harbored pCDF-DUETΩ*cas1-cas2-csn2* and pEC85ΩtracrRNA-Leader-CRISPR. The bacterial culture was sub-grown daily for two weeks and the spacer acquisition activity was monitored. If these three *cas* genes are sufficient for spacer acquisition, prespacers would be sampled from the plasmids or from the genome (Wei et al., 2015a) and incorporated into the type II-A CRISPR array on the plasmid. To analyze the spacer acquisition activity, an aliquot of the bacterial culture was sampled daily for total DNA extraction, and the extracted DNA was used as PCR template for amplifying the CRISPR arrays of *S. pyogenes* and *E. coli,* respectively. The parental DNA band (PCR-amplified CRISPR array without spacer uptake) is expected to expand 65-66-nt in size with every single spacer being acquired (**Figure 7A**).

There are two CRISPR arrays present in the genome of *E. coli* BL21-AI, one of which (array I, type I-E systems) contains a conserved leader sequence (Yosef et al., 2012). A type I-E plasmid-based acquisition study in *E. coli* BL21-AI detected spacer acquisition only in the CRISPR array I with a conserved leader (Yosef et al., 2012). We wondered whether type II-A system is active in *E. coli*. Therefore, the CRISPR array I of *E. coli* BL21-AI was also monitored for spacer acquisition. Our result showed that expression of *cas1*, *cas2*, *csn2* and tracrRNA from the type II-A system in *E. coli* BL21-AI did not confer new spacers in the CRISPR array I of *E. coli* (data not shown) (**Supplementary Table S1**). Pairwise sequence alignment of the leader and the repeat sequences of the type II-A system of *S. pyogenes* and the type I-E system of *E. coli* BL21-AI showed low sequence identities (**Supplementary Figure S1**). This explains that it is unlikely for type II-A adaptation machinery to integrate spacer into the type I-E CRISPR array of *E. coli* BL21-AI. In line with this, type II-A system relies on LAS sequence to direct Cas1-Cas2 towards the spacer integration site (McGinn and Marraffini, 2016), whereas type I-E system is dependent on IHF (Nunez et al., 2016). PCR monitoring showed that spacer acquisition was also not detected in the *S. pyogenes* CRISPR array on the plasmid (**Figure 7B; Supplementary Table S1**).

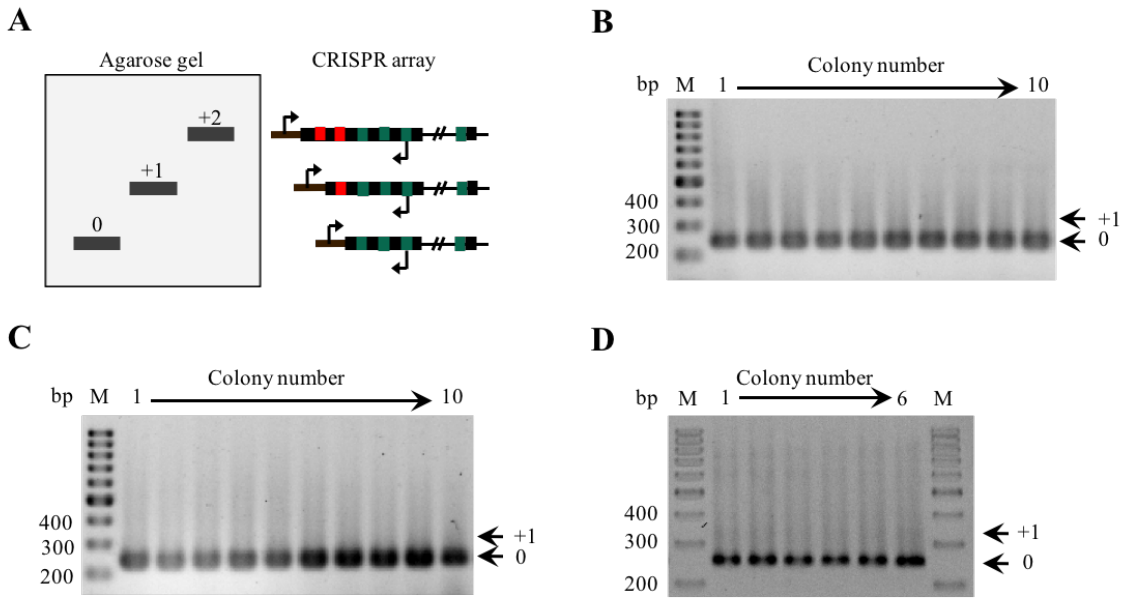


Figure 7. The spacer acquisition screening of *S. pyogenes* heterologous system in *E. coli* BL21-AI.

(A) Schematics of PCR screening for acquisition. A single spacer acquisition gives a 65-66-nt increment of size to the parental band. Arrows indicate forward and reversed primers. '0', parental band without acquisition; '+1', DNA band with one newly acquired spacer; '+2', DNA band with two newly acquired spacers; black rectangles, repeats; green rectangles, existing spacers; red rectangles, newly acquired spacers. **(B)** Spacer acquisition screening for *S. pyogenes* heterologous system challenged with plasmids. *E. coli* BL21-AI harboring two plasmids, *i.e.* pCDF-DUETΩ*cas1-cas2-csn2* and pEC85ΩtracrRNA-Leader-CRISPR was used in this assay. No expanded band (no spacer acquisition) was detected in this assay. **(C)** Spacer acquisition screening for *S. pyogenes* heterologous system challenged with plasmids. *E. coli* BL21-AI harboring three plasmids, *i.e.* pCDF-DUETΩ*cas1-cas2-csn2*, pEC85ΩtracrRNA-Leader-CRISPR-*cas9* and pUC19ΩSpy_0700-NTG(PAM) was studied in this assay. Expanded band was not observed in this assay. **(D)** Spacer acquisition screening for *S. pyogenes* heterologous system challenged with lytic phage Lambda. *E. coli* BL21-AI harboring two plasmids, *i.e.* pCDF-DUETΩ*cas1-cas2-csn2* and pEC85ΩtracrRNA-Leader-CRISPR-*cas9* was challenged with phage Lambda in this assay. The agarose gel picture shows the acquisition result of the phage challenge with MOI of 1. This assay did not detect any expanded band. Representative 2% agarose gel pictures for the spacer acquisition screenings are shown here. '0', the expected size of the parental band without acquisition; '+1', the expected size of the DNA band with one newly acquired spacer. M, GeneRuler™ 100-bp DNA Ladder (Thermo Scientific).

We did not detect spacer acquisition in the heterologous system expressing *cas1, cas2* and *csn2* and tracrRNA, and we questioned whether it was due to the lack of Cas9. To investigate whether Cas9 is required for spacer acquisition in the type II-A systems, we included Cas9 in the plasmid-based spacer acquisition assay, by introducing pEC85ΩtracrRNA-Leader-CRISPR-*cas9*, of which the expression of *cas9* was controlled under the native promoter of the *cas* operon. Since it is unclear whether primed spacer acquisition exists in the type II-A systems, we wondered if Cas9 plays a similar role like the interference machinery in the primed spacer

acquisition of the type I-E systems (Datsenko et al., 2012; Swarts et al., 2012). Therefore, we designed a three-plasmid system that mimics the primed acquisition by including pCDF-DUETΩ*cas1-cas2-csn2*, pEC85ΩtracrRNA-Leader-CRISPR-*cas9* and pUC19ΩSpy_0700-NTG(PAM). Spy_0700-NTG(PAM) is a protospacer with a mutated PAM, (the NGG PAM was mutated to NTG), which is supposed to accelerate the spacer acquisition activity via priming if primed spacer acquisition exists in the type II-A system. The three-plasmid acquisition assay was similar to the two-plasmid acquisition assay described earlier, with a small modification for the sampling of the template DNA for PCR screening (see Materials and Methods). In this assay, the colonies that had lost pUC19ΩSpy_0700-NTG(PAM) were selected via replica plating on the non-selective and selective plates, followed by colony PCR screening. Spacer acquisition was not detected in *S. pyogenes* CRISPR array, as well as the *E. coli* CRISPR array (**Figure 7C; Supplementary Table S1**). In summary, spacer acquisition was not detected in the two- and three-plasmid acquisition assay.

3.1.1.2 Spacer acquisition in the *S. pyogenes* heterologous system with phage challenge

The plasmid-based acquisition was possibly not able to confer strong selective pressure to the bacteria to activate spacer acquisition. To trigger spacer acquisition with high selective pressure and positively select the bacteria that acquire and survive, we challenged the bacteria with a virulent variant of phage Lambda (phage λ_{vir}) (Brouns et al., 2008), which lacks its lysogenic regulatory regions. In this study, a two-plasmid system comprised of pCDF-DUETΩ*cas1-cas2-csn2* and pEC85ΩtracrRNA-Leader-CRISPR-*cas9* was used. Since spacer acquisition would eventually lead to interference while the cells are challenged by a cognate invader, wild type (WT) *cas9* is crucial for the phage challenge assay to ensure the survival of the cells, unless they were protected by other anti-viral defense systems. The bacteria that survived the phage challenge would appear as colonies on the plates, and these colonies were screened for spacer acquisition with colony PCR. Various multiplicities of infections (MOIs), *i.e.* MOIs of 1, 10 and 100 have been used for the phage infection, yet, no spacer acquisition was detected (**Figure 7D; Supplementary Table S1**).

Altogether, we did not detect spacer acquisition in the *S. pyogenes* heterologous system. Since spacer acquisition of the type II-A system has not been studied heterologously in the *E. coli*,

we did not know whether the *S. pyogenes* heterologous system is active in *E. coli*. Hence, we also studied spacer acquisition of the type II-A systems in *S. pyogenes*.

3.1.2 Plasmid challenge spacer acquisition study in *S. pyogenes* SF370

Although the heterologous system is more convenient for genetic manipulation, it is unsuitable for studying the potential strain-specific host factor(s) necessary for spacer acquisition, which is better to study in the native host. To investigate endogenous spacer acquisition, the shuttle vector pEC85, which serves as a prespacer source, was transformed into the *S. pyogenes* SF370 WT strain. The WT strain was grown in THY medium without antibiotic to allow the occurrence of spacer acquisition from the plasmid without killing the cells that have lost the plasmid. The colonies from every cycle of culture were checked for spacer acquisition by PCR. Every cycle of sub-growing represents every time the culture was transferred to a fresh culture for growing. Spacer acquisition was monitored until the 30th cycle, yet no acquisition was detected under the tested condition (data not shown) (**Supplementary Table S1**).

In the WT strain, spacer acquisition will lead to the cleavage of the plasmid by Cas9, and the use of antibiotic for the maintenance of the plasmid could result in cell death. To maintain the survival of the acquisition-positive cells and improve the plasmid maintenance with the use of antibiotic, a *cas9* deletion mutant was complemented with pEC85 harboring either tracrRNA-*cas9-D10A*-speM (a Cas9 RuvC mutant) or tracrRNA-*cas9-H840A*-speM (a Cas9 HNH mutant). The mutation of the catalytic residue in the RuvC or HNH domain of Cas9 allows Cas9 mutants to nick the plasmid DNA instead of cleaving the dsDNA (Jinek et al., 2012), and it would not disturb spacer acquisition (Heler et al., 2015; Wei et al., 2015a). Here, we included speM, a protospacer that matches a spacer in the type II-A CRISPR array of *S. pyogenes*, to check whether this could accelerate the spacer acquisition activity in a manner similar to the primed spacer acquisition in the type I-E systems (Datsenko et al., 2012; Swarts et al., 2012). In this study, no spacer acquisition was detected until the 56th day of sub-culturing period for all the *S. pyogenes cas9* deletion mutant harboring pEC85 vector or pEC85ΩspeM plasmid, as well as Cas9 RuvC mutant with tracrRNA and speM, and Cas9 HNH mutant with tracrRNA and speM (**Figure 8; Supplementary Table S1**). Altogether, we did not detect spacer acquisition in the heterologous type II-A system of *S. pyogenes*, as well as the native type II-A system *S. pyogenes*.

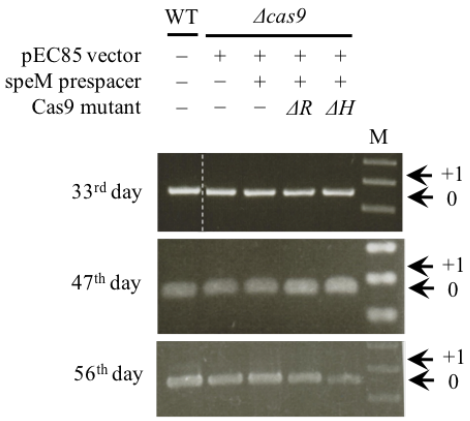


Figure 8. The screening for the plasmid-based spacer acquisition of endogenous type II-A system of *S. pyogenes*.

S. pyogenes Δcas9 was transformed with pEC85 vector, pEC85ΩspeM, pEC85ΩtracrRNA-*cas9-D10A*-speM (a Cas9 RuvC mutant) (abbreviated as *ΔR*) or pEC85ΩtracrRNA-*cas9-H840A*-speM (a Cas9 HNH mutant) (abbreviated as *ΔH*) and monitored for spacer acquisition via PCR screening. Selected agarose gel (1.5%) pictures from the 33rd, 47th and 56th day of sub-culturing are indicated here. '0', the expected size of the parental band without acquisition; '+1', the expected size of the DNA band with one newly acquired spacer. M, GeneRuler™ 100-bp DNA Ladder (Thermo Scientific).

3.2 Spacer acquisition in the type II-A CRISPR-Cas system of *S. thermophilus* LMD-9

Due to unsuccessful detection of *S. pyogenes* spacer acquisition activity *in vivo*, we decided to use *S. thermophilus* as a model organism for the investigation of type II-A spacer acquisition. *S. thermophilus* has been shown to be naturally active in spacer acquisition (Barrangou et al., 2007; Horvath et al., 2008), and the type II-A CRISPR-Cas systems of *S. thermophilus* LMD-9 and *S. pyogenes* SF370 are similar (**Supplementary Table S2**) (Chylinski et al., 2014; Fonfara et al., 2014; Makarova et al., 2011).

3.2.1 The heterologous type II-A CRISPR-Cas system of *S. thermophilus* is established in *E. coli* BL21-AI

Similar to *S. pyogenes*, the type II-A system of *S. thermophilus* LMD-9 was studied in *E. coli* BL21-AI to obtain insights in the acquisition mechanism. Our study focused on CRISPR1 (Cr1) locus of the strain LMD-9, as the CRISPR1 locus of

S. thermophilus DGCC7710 was previously described for being more active in spacer acquisition than CRISPR3 (Horvath et al., 2008). Pairwise sequence alignment showed that the Cas1 from the CRISPR1 loci of LMD-9 and DGCC7710 strains are 100% identical, whereas their Cas2, Csn2 and Cas9 are also almost identical (**Supplementary Table S3; Supplementary Figure S2**), which is in agreement with the literature (Chylinski et al., 2014; Fonfara et al., 2014). Unless otherwise specified, all the *cas* genes or Cas proteins mentioned here refer to those from CRISPR1 locus. The heterologous system of *S. thermophilus* was established in a similar way as the heterologous system of *S. pyogenes*. The transcription of *csn2* gene of the pCDF-DUETΩ*cas1-cas2-csn*2 was confirmed by semi-quantitative RT-PCR, which could represent the transcription of the whole *cas* operon as described before (**Figure 9**). The proteins expression was not verified by Western Blot analysis due to the absence of the antibodies against *S. thermophilus* Cas proteins at that time.

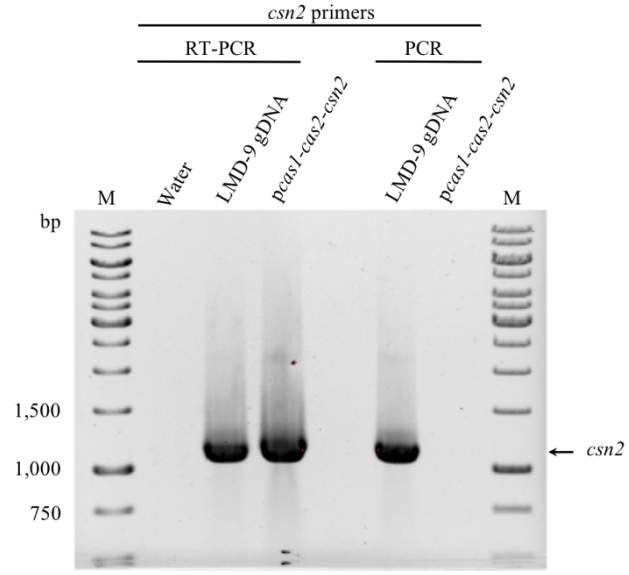


Figure 9. Verification of the transcription of *csn2* in the heterologous type II-A CRISPR-Cas system of *S. thermophilus* in *E. coli* BL21-AI.

RNA expression of heterologous *S. thermophilus* type II-A CRISPR-Cas system in *E. coli* BL21-AI was verified with semi-quantitative RT-PCR. 2% agarose gel picture of RT-PCR showing the expression level of *csn2* gene of pCDF-DUETΩ*cas1-cas2-csn*2. The *cas* genes expression was induced in *E. coli* BL21-AI with 1 mM IPTG and 0.2% arabinose for 2 hours. A PCR reaction was additionally performed to confirm that the RNA templates were free from DNA contamination. pCDF-DUET is abbreviated as "p" in the picture. LMD-9 gDNA, genomic DNA of *S. thermophilus* LMD-9; M, GeneRuler™ 1-kb DNA Ladder (Thermo Scientific).

3.2.1.1 The plasmid-based spacer acquisition in the heterologous system of *S. thermophilus*

We started the plasmid-based acquisition study with *E. coli* harboring pCDF-DUETΩ*cas1-cas2-csn*2 and pEC85ΩLeader-CRISPR1. These two heterologous plasmids served as prespacer sources due to the distribution of a number of PAMs, NNAGAAW (Horvath et al., 2008), throughout the plasmids. This study was similar to the *S. pyogenes* heterologous system with some modifications (refer to Materials and Methods for details). The induced culture was monitored from the 1st cycle until the 24th cycle of sub-growing, however, no acquisition was detected (data not shown). In addition to that, we also monitored the spacer acquisition from the culture that had lost pCDF-DUETΩ*cas1-cas2-csn*2 with the presumption that plasmid loss might be caused by acquisition. For this assay, an aliquot of the culture from the 16th cycle was plated on a LB (lysogeny broth) plate with non-selective antibiotic for pCDF-DUETΩ*cas1-cas2-csn*2, followed by replica plating on the non-selective and selective plates for pCDF-DUETΩ*cas1-cas2-csn*2. Only 4 colonies out of 200 colonies grew on the selective plate for pCDF-DUETΩ*cas1-cas2-csn*2, which indicated that most of the colonies had lost pCDF-DUETΩ*cas1-cas2-csn*2. For those colonies that have lost pCDF-DUETΩ*cas1-cas2-csn2*, 38 colonies were screened for spacer acquisition, however, spacer acquisition was not detected (**Figure 10A**).

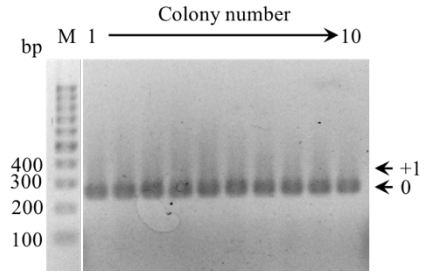
3.2.1.2 Phage challenge spacer acquisition in the heterologous system of *S. thermophilus*

To trigger spacer acquisition by introducing higher selective pressure, we used phage λ_{vir} to challenge the *S. thermophilus* heterologous system containing pCDF-DUETΩ*cas1-cas2-csn2* and pEC85ΩLeader-CRISPR1. To investigate the optimal infection conditions for triggering the acquisition activity, various MOIs and adsorption times (bacteria-phage incubation time at 37°C without agitation) were tested. This included MOIs of 0.1, 1 and 10 with adsorption times of 5, 10, 15 and 20 minutes, respectively. Despite trying numerous phage challenge conditions, none of the screened colonies were positive for spacer acquisition (**Figure 10B**).

In conclusion, we did not detect spacer acquisition in the heterologous system that harbored *cas1*, *cas2* and *csn2*. While the acquisition study with *cas1*, *cas2* and *csn2* was performed, the pEC85ΩCRISPR1-tracrRNA-Cas9 was still in process of cloning. Later, two studies showed that Cas9 and tracrRNA are needed for spacer acquisition (Heler et al., 2015; Wei et al., 2015a).

The absence of Cas9 and tracrRNA in our heterologous system is a reason that spacer acquisition was not detected.

A



B

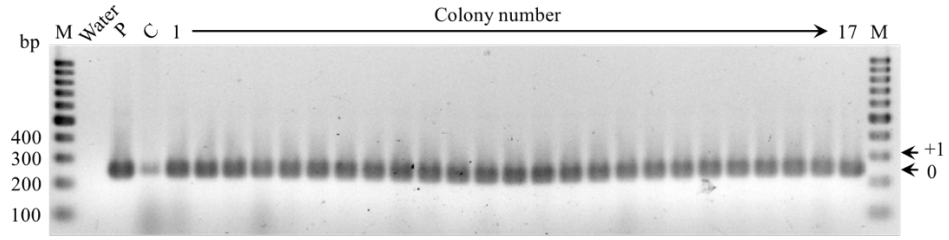


Figure 10. The spacer acquisition screening of *S. thermophilus* heterologous system in *E. coli* BL21-AI.
(A) Spacer acquisition screening for *S. thermophilus* heterologous system challenged with plasmid. *E. coli* BL21-AI harboring two plasmids, *i.e.* pCDF-DUETΩ*cas1-cas2-csn2* and pEC85ΩLeader-CRISPR1, was studied in this assay. This agarose gel picture shows the acquisition screening of the colonies from the cycle 16, which had lost pCDF-DUETΩ*cas1-cas2-csn*2. No expanded band (no spacer acquisition) was detected in this assay. A representative 2% agarose gel picture for the spacer acquisition screenings is shown here. **(B)** Spacer acquisition screening for *S. thermophilus* heterologous system challenged with lytic phage Lambda. *E. coli* BL21-AI harboring two plasmids, *i.e.* pCDF-DUETΩ*cas1-cas2-csn2* and pEC85ΩLeader-CRISPR1, was studied in this assay. Expanded band was not detected in this assay. A representative 2% agarose gel for the screenings is shown here. Controls for the acquisition study are shown in lane 1 to 3, where 'P' indicates pEC85ΩCRISPR1, and 'C' indicates colony harboring this plasmid. '0', the expected size of the parental band without acquisition; '+1', the expected size of the DNA band with one newly acquired spacer. M, GeneRuler™ 100-bp DNA Ladder (Thermo Scientific).

3.2.2 The endogenous type II-A system of *S. thermophilus* LMD-9 is active in spacer acquisition

3.2.2.1 Phage challenge in *S. thermophilus* shows active spacer acquisition

We started the spacer acquisition study with the heterologous system, as we did not have the lytic phages for *S. pyogenes* and *S. thermophilus* at the beginning of our studies. As soon as the lytic phage (phage DT1) for *S. thermophilus* LMD-9 was available, we challenged *S. thermophilus* with phage DT1 to study spacer acquisition. The bacteria that survived from phage challenge were screened for spacer acquisition via colony PCR. The WT strain showed acquisition in both CRISPR1 and CRISPR3 loci, with higher spacer acquisition rate in the CRISPR3 locus compared to the CRISPR1 locus (**Figure 11; Table 1**). Spacer acquisition of a double knock-out mutant of *S. thermophilus* LMD-9, with the deletions of the entire CRISPR2 and CRISPR3 loci (known as ΔCr2ΔCr3 strain), was compared to the WT in order to test for a potential increase of the spacer acquisition rate in this mutant. Our results, however, showed no significant difference in the rate of spacer acquisition in the CRISPR1 loci between the WT and the ΔCr2ΔCr3 strains (**Table 1**). In summary, spacer acquisition was detected in both CRISPR1 and CRISPR3 loci when *S. thermophilus* LMD-9 WT was challenged with phage DT1.

Table 1. The number of colonies with newly acquired spacers upon phage challenge in *S. thermophilus* LMD-9 WT and ΔCr2ΔCr3 mutant.

Strains	CRISPR1	CRISPR3
WT	8% (11/131 colonies)	45% (55/130 colonies)
ΔCr2ΔCr3	8% (4/51 colonies)	-

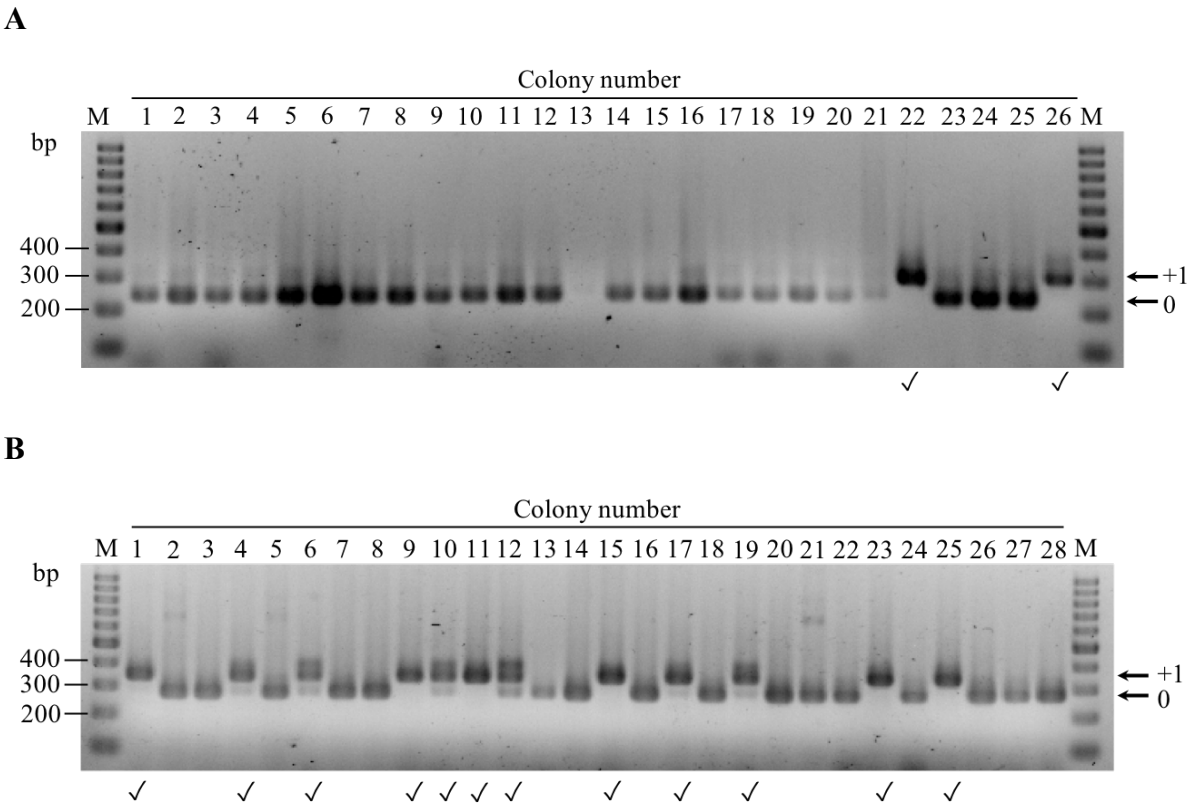


Figure 11. Both CRISPR1 and CRISPR3 loci of *S. thermophilus* LMD-9 WT showed spacer acquisition upon phage challenge.

A representative result of endogenous *S. thermophilus* spacer acquisition study with phage challenge for **(A)** CRISPR1 and **(B)** CRISPR3 loci. *S. thermophilus* WT strain was infected with phage DT1 at MOIs of 1 or 10. The survivors of the phage challenge were screened for spacer acquisition with colony PCR and analyzed on 2% agarose gel. '0', the expected size of the parental band without acquisition; '+1', the expected size of the DNA band with one newly acquired spacer; ✓, new spacer being acquired; M, GeneRuler™ 100-bp DNA Ladder (Thermo Scientific).

3.2.2.2 Cas proteins over-expression in *S. thermophilus* increases spacer acquisition

To allows easy genetic manipulation of *cas* genes, we investigated spacer acquisition in *S. thermophilus* LMD-9 by over-expressing the Cas proteins. We did not have a suitable vector for *S. thermophilus*, therefore we examined whether the spacer acquisition in LMD-9 could be facilitated by over-expressing the CRISPR1 Cas proteins of DGCC7710 in LMD-9 WT, as the CRISPR1 loci of *S. thermophilus* LMD-9 and DGCC7710 strains are close orthologs (**Supplementary Table S3; Supplementary Figure S2**) (Chylinski et al., 2014; Fonfara et al., 2014). Elevating the expression of Cas9, Cas1, Cas2 and Csn2 increases the spacer acquisition

rate in the CRISPR1 array of LMD-9 WT and ΔCr2ΔCr3 strains (**Figure 12A**), which is in agreement with the literature (Wei et al., 2015a). Nevertheless, increasing the expression of all the four Cas proteins from the CRISPR1 locus, did not increase the spacer acquisition in the CRISPR3 locus of the WT strain, suggesting a lack of crosstalk between the Cas proteins from CRISPR1 and CRISPR3 loci (**Figure 12B**). We further confirmed that Cas9 is essential for spacer acquisition as spacer acquisition was not detected in ΔCr2ΔCr3Δ*cas9_Cr1* (a triple knock-out mutant, *i.e.* deletion of *cas9* from the CRISPR1 locus of the ΔCr2ΔCr3 strain) when only Cas1, Cas2 and Csn2 of the CRISPR1 were over-expressed (**Figure 12A**), which is line with other studies (Heler et al., 2015; Wei et al., 2015a). Altogether, we showed that spacer acquisition in the type II-A systems is locus-specific, and we also confirmed the requirement of all the Cas proteins in spacer acquisition (Heler et al., 2015; Wei et al., 2015a), which led us to further investigate the interactions among the Cas proteins.

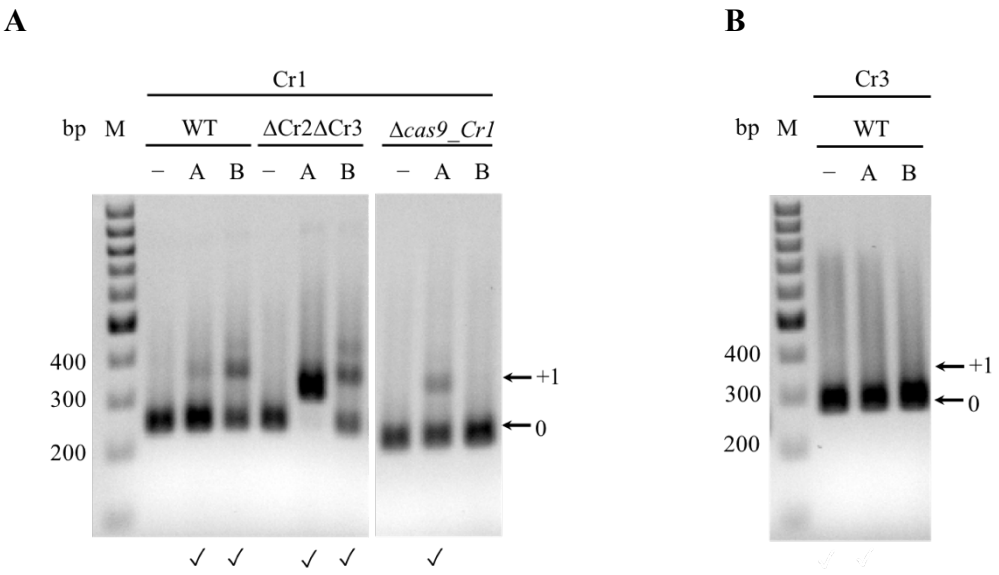


Figure 12. Over-expression of *cas* genes in *S. thermophilus* LMD-9 increases spacer acquisition.
cas1cas2csn2cas9 or *cas1cas2csn2* of the CRISPR1 locus of *S. thermophilus* DGCC7710 strain were over-expressed in the close ortholog, *S. thermophilus* LMD-9. Spacer acquisition on **(A)** CRISPR1 (Cr1) and **(B)** CRISPR3 (Cr3) arrays was examined via PCR. PCR reactions were visualized on 2% agarose gel. ΔCr2ΔCr3, a CRISPR2 and CRISPR3 double knock-out mutant of LMD-9; ΔCr2ΔCr3Δ*cas9*_Cr1, a CRISPR2, CRISPR3 and *cas9* (CRISPR1) triple knock-out mutant of LMD-9. ΔCr2ΔCr3Δ*cas9*_Cr1 is abbreviated as Δ*cas9*_Cr1 in this figure. '–', without plasmid; 'A', with plasmid p*cas1-cas2-csn2-cas9*; 'B', with plasmid p*cas1-cas2-csn2*; '0', parental band without acquisition; '+1', DNA band with one newly acquired spacer; ✓, new spacer being acquired, M, GeneRuler™ 100-bp DNA Ladder (Thermo Scientific).

3.3 Characterization of protein-protein interactions of type II-A CRISPR-Cas systems

3.3.1 Cas proteins interact with proteins within and beyond CRISPR-Cas systems

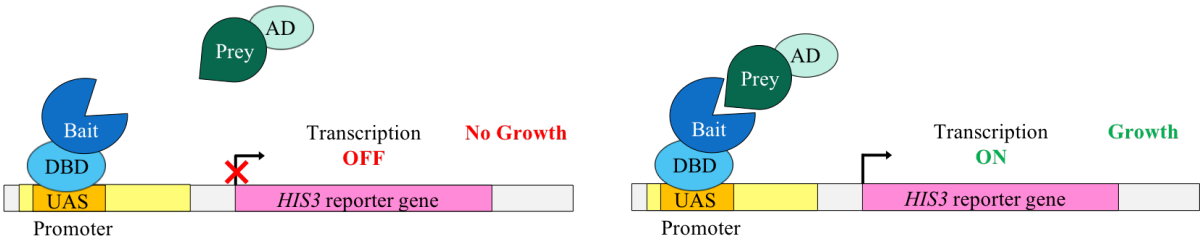
Protein-protein interactions (PPIs) within the Cas family during various stages of CRISPR immunity have been reported. For instance, complex formations were reported for Cas1 and Cas2 in type I-E and II-A systems (Nunez et al., 2014; Xiao et al., 2017), Cas1 and Cas2-Cas3 fusion protein in type I-F system of *P. atrosepticum* (Fagerlund et al., 2017), Cas1 and Cas4 in type I-C of *B. halodurans* (Lee et al., 2018). Besides, Cas1 interacts with Cas6 and Cas7 subunits of the Cascade in the type I-E system, as well as DNA repair proteins such as RecB, RecC and RuvB (Babu et al., 2011). In the type II-A systems of *S. pyogenes*, Cas9, Cas1, Cas2 and Csn2 were co-purified when N-terminal His_6-tagged Cas9 was used as a bait in Ni-NTA affinity chromatography, and when C-terminal His_6-tagged Csn2 was used as a bait in ion exchange chromatography (Heler et al., 2015). Altogether, these studies showed CRISPR-Cas type-specific interactions between various Cas proteins for different functions, as well as their interactions with proteins beyond CRISPR-Cas systems.

Although all Cas proteins of the type II-A systems, *i.e.* Cas1, Cas2, Csn2 and Cas9, are essential for spacer acquisition (Heler et al., 2015; Wei et al., 2015a), the details of their interactions with one another are still obscure. We hypothesized that the Cas1-Cas2 complex acts as the core proteins that only interact with Cas9 and Csn2 during a specific mechanistic step of spacer acquisition. Studies have revealed the involvement of the non-CRISPR proteins in the type I-E primed spacer acquisition (Ivancic-Bace et al., 2015), as well as the connections of the DNA repair pathways with several CRISPR-Cas systems (Babu et al., 2011; Klaiman et al., 2014; Levy et al., 2015; Modell et al., 2017; Williams et al., 2007). Hence, we speculated that type II-A Cas proteins also interact with other host factors, especially during the spacer acquisition stage. We examined these hypotheses by identifying the interacting partners for the type II-A Cas proteins and investigated the interactions regions.

3.3.1.1 Yeast two-hybrid screening for the interacting partners of *S. pyogenes* Cas proteins

We hypothesised that type II-A Cas proteins interact among themselves and also with other host proteins for different purposes. Therefore, we used the yeast two-hybrid (Y2H) screen service provided by Hybrigenics Services to study the PPIs of *S. pyogenes* Cas proteins, as we were studying *S. pyogenes* spacer acquisition at that time. The Y2H screen is a high throughput technique for detecting PPIs via examining the physical interactions of two proteins in yeast cells. This technique allows the positive selection of yeast in a histidine-free medium, when a protein of interest (bait) that is fused to a DNA binding domain (DBD) interacts with a potential interacting partner (prey) that is fused to an activation domain of a transcription factor, resulting in the transcription of *HIS3* reporter gene and histidine synthesis (**Figure 13A**) (Brückner et al., 2009; Stasi et al., 2015; Westermarck et al., 2013).

A



B

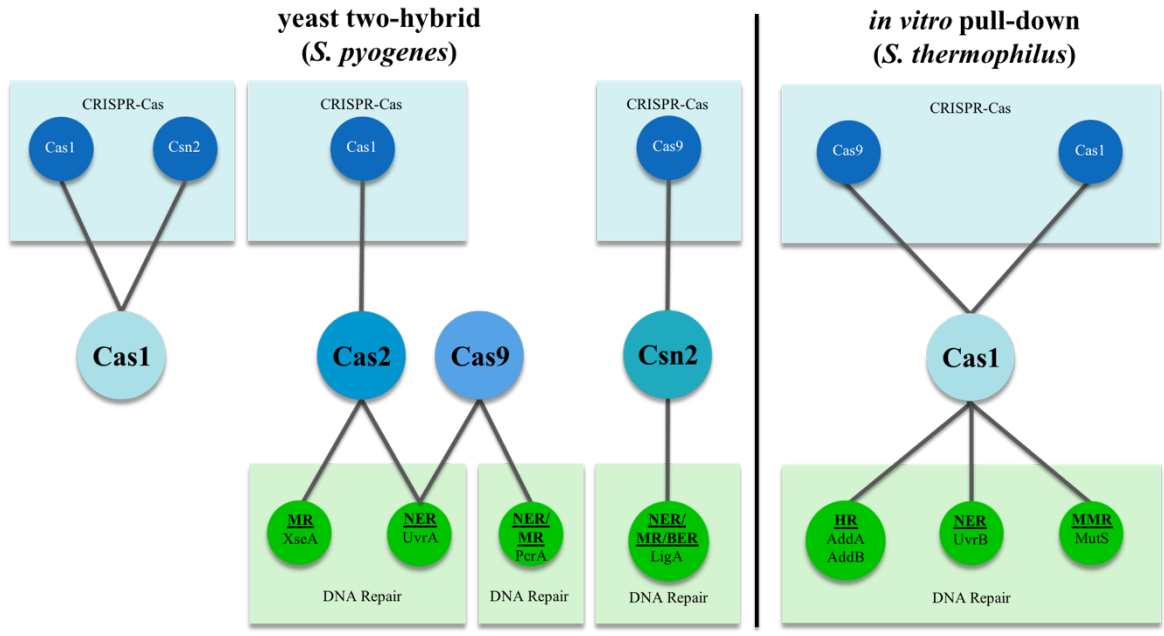


Figure 13. Cas proteins interact with proteins from different pathways.
(A) Schematics of yeast two-hybrid systems. A bait and a potential prey are respectively fused to a DNA binding domain (DBD) and an activation domain (AD) of a transcription factor of the *HIS3* reporter gene. The bait-DBD binds to the upstream activator sequence (UAS) of the promoter. (Left panel) In the absence of the interaction between bait and prey, there is no transcription of *HIS3*. Therefore, yeast cannot grow on the histidine-free medium. (Right panel) The interaction of the prey with the bait brings the activation domain of the transcription factor to the proximity of the DNA binding domain of the promoter. This allows the transcription of *HIS3* and synthesis of histidine. Consequently, yeast can grow on the histidine-free medium. **(B)** Illustration summarizes the Cas proteins' interacting partners from the CRISPR-Cas systems (CRISPR1) and the DNA repair pathways, detected by the yeast two-hybrid screening (*S. pyogenes* SF370) and the *in vitro* pull-down (*S. thermophilus* LMD-9). MMR, mismatch repair; NER, nucleotide excision repair; BER, base excision repair; HR, homologous recombination.

The interaction reliability in Y2H is evaluated by Predicted Biological Score (PBS), with values ranging from A to F, which are the thresholds that are defined by the probability of an interaction to be non-specific via comparing the amount of independent prey fragments detected versus the background (their random detection). The scores of A, B and C indicate that the interaction is very highly confident, highly confident or confident, respectively. The PBS of D shows that the interaction is moderately confident, which comprises a mix of false positives or hardly detectable interaction that is caused by under-representation of the mRNA in the library, prey folding or prey toxicity in yeast. Therefore, careful verification is needed for the PBS of D. Risk of unspecific interaction and experimentally proven technical artifacts are labeled as PBS of E and F, respectively (Rain et al., 2001).

In this study, either Cas1, Cas2, Csn2 or Cas9 were used as a bait for screening a cDNA library of potential preys that are cloned into plasmids to determine their interacting partners. Y2H analysis revealed that the interacting partners of the Cas proteins are mainly the proteins from the CRISPR-Cas systems and DNA repair/homologous recombination (HR) (**Figure 13B; Supplementary Tables S4-S7**), where some of interacting partners share the same pathways as the interacting partners of Cas1 identified via the *in vitro* pull-down assay (see section 3.3.1.2) or correspond to other studies (Babu et al., 2011; Kim et al., 2013; Wiedenheft et al., 2009; Xiao et al., 2017). Among the candidates, the Cas1-Cas1 (prey is underlined) (PBS of A) (**Supplementary Table S4**) and Cas2-Cas1 (PBS of B) (**Supplementary Table S5**) interactions are in agreement with Cas1 dimer and Cas1-Cas2 complex formation reported in the literature (Nunez et al., 2014; Xiao et al., 2017). The Cas1-Csn2 interaction (PBS of C)

(**Supplementary Table S4**) indicated in this study conforms to SEC experiments that provided the evidence of interaction of these two proteins (Ka et al., 2016). Our Y2H study additionally revealed the direct interaction of Csn2-Cas9 (PBS of D) (**Supplementary Table S6**), which was also confirmed by a very recent study (Ka et al., 2018). The PBS of D obtained by Csn2-Cas9 interaction might be due to the toxicity of Cas9 in yeast, which was observed in the Y2H analysis that used Cas9 as a bait.

Our Y2H screen also demonstrated for the first time the interactions between several Cas proteins and proteins from three DNA repair pathways, which are the mismatch repair (MMR), nucleotide excision repair (NER) and base excision repair (BER) pathways (**Figure 13B; Supplementary Table S5-S7**). For instance, Cas2 interacts with XseA from MMR and UvrA from NER pathway (PBS of A for both candidates) (**Supplementary Table S5**); Cas9 interacts with UvrA from NER (PBS of A) and PcrA from NER and MMR pathways (PBS of B) (**Supplementary Table S7**); and Csn2 interacts with LigA from NER, MMR and BER (PBS of A) (**Supplementary Table S6**). Among these candidates, the interactions of Cas2-UvrA and Cas9-UvrA were suggested by PBS of A as very highly reliable interactions. Notably, the interactions of Cas proteins and the proteins from NER pathway are also supported by *in vitro* pull-down experiment (see section 3.3.1.2). Previous study in type I-E system of *E. coli* revealed the interactions of Cas1 and the DNA repair proteins (Babu et al., 2011). Therefore, it is not unlikely that type II-A Cas proteins interact with DNA repair proteins, although more studies are needed to confirm that.

Altogether, our Y2H unveils the direct interactions among the Cas proteins, such as Cas2-Cas1, Cas1-Csn2 and Csn2-Cas9; as well as the interactions with DNA repair proteins, such as Cas2-XseA, Cas2-UvrA, Cas9-UvrA, Cas9-PcrA and Csn2-LigA. The Y2H results for *S. pyogenes* SF370 were obtained after we started using *S. thermophilus* LMD-9 for studying type II-A spacer acquisition. Since the type II-A CRISPR-Cas systems of both *S. pyogenes* and *S. thermophilus* are similar (**Supplementary Table S2**) (Barrangou and van der Oost, 2013; Makarova et al., 2011), comparable results were expected by using the *in vitro* pull-down assay in *S. thermophilus*.

3.3.1.2 *In vitro* pull-down of *S. thermophilus* Cas1 revealed interacting partners from various pathways

Unlike Y2H, pull-down allows the discovery of novel interacting proteins in their native cellular environment, thereby it enables the detection of the interacting proteins that might be regarded as false negatives in Y2H due to the lack of native cellular environment (e.g. cytosolic or membrane-bound protein) for the protein interactions. *In vitro* pull-down assay is based on the principle that a purified tagged protein (bait) is immobilized on a resin to capture and "pulls-down" the interacting protein (prey) from the cell lysates after several washes to remove the unspecifically bound proteins. The eluted bait-prey proteins are analyzed on SDS-PAGE gel and subsequently analyzed by mass-spectrometry. Pull-down assay is a technique to investigate physical protein interactions between two or more proteins with strong and stable interactions, but it is challenging to study transient protein interactions, which may dissociate during the pull-down experiment (Brückner et al., 2009).

The Y2H screen described above revealed the interacting partners for the type II-A Cas proteins of *S. pyogenes*. After changing the model organism to *S. thermophilus* LMD-9, an *in vitro* pull-down assay was used as an alternative approach to verify the interacting partners of the type II-A Cas1 (CRISPR1) and further investigate the PPIs of Cas1. We decided to focus on Cas1, because it is one of the core proteins of the adaptation machinery that executes numerous key mechanistic steps during spacer acquisition. In this pull-down assay, we used C-terminal Cysteine Protease Domain (CPD)-His_{12}-tagged Cas1 as a bait to pull-down the interacting partners from the cell lysate of *S. thermophilus* LMD-9 WT strain, because a cleaner background could be obtained via cleaving the CPD-His_{12}-tag while eluting the bait-prey proteins from the affinity column. Here, the negative controls were: (1) The cell lysate only control, which serves to identify and exclude the false positives, *i.e.* proteins that are unspecifically bound to the resin; and (2) the Cas1 bait only control, which helps to identify and exclude the false positives caused by proteins that are unspecifically bound to the tag of the bait. Cas1 bait only control also additionally serves as a positive control to confirm the binding of the tagged-bait to the resin.

To obtain the fold change of the pull-downed interacting partners, a software named Scaffold was used to compare the mass-spectrometry results of the bait-prey sample with the controls. Here, the result of the bait-prey sample containing the interacting proteins was used to subtract the one of the cell lysate only control, and the obtained result was subsequently compared to

the bait only control to remove the background. While a fold change with a cutoff of ≥2 is widely used, a more stringent cutoff of ≥3 was used here to narrow down the numbers of the candidates that we were interested in. With a fold change cutoff of ≥3, the interacting partners of Cas1 identified from the mass-spectrometry analysis included Cas proteins, DNA repair proteins, ATP-binding cassette transporters (ABC transporters) and DNA replication proteins (**Supplementary Table S8**). Apart from the ABC transporters and the DNA replication proteins, most of the identified interacting partners correspond to the literature either directly (*i.e.* direct physical interactions) or indirectly (*i.e.* the proteins are involved in the same mechanistic steps, such as Cas1 and AddAB (Modell et al., 2017)), as shown in the (**Figure 13B; Table 2**). Therefore, the candidates from the CRISPR-Cas systems and DNA repair pathways seem to be more promising and they are our main focus here.

Table 2. Selected interacting partners of *S. thermophilus* Cas1 (CRISPR1) obtained from *in vitro* pull-down assay and confirmed by either yeast two-hybrid (Y2H) or literature.

System/ Pathway	Stage/pathway	Interacting partners	Fold change[a]	Literature
CRISPR-Cas	Adaptation; Interference	Cas9	8.8	(Heler et al, 2015[b]; Wei et al, 2015[b])
	Adaptation	Cas1	4.1	(Babu et al, 2011; Kim et al, 2013; Wiedenheft et al, 2009; Ka et al, 2016; Xiao et al, 2017)
DNA repair	Homologous recombination (HR)	AddB	3.4	(Babu et al, 2011; Levy et al, 2015[b]; Modell et al, 2017[b])
		AddA	3.1	(Babu et al, 2011; Levy et al, 2015[b]; Modell et al, 2017[b])
	Nucleotide excision repair (NER)	UvrB	3.5	Correlate with UvrA (candidate of Y2H-Cas2 and Y2H-Cas9)
	Mismatch repair (MMR)	MutS	3.6	(Babu et al, 2011)

[a] Only candidates with fold change ≥3.0 are shown

[b] The direct physical interaction between Cas1 and the candidate has not been experimentally shown. However, Cas1 and the candidate are correlated in the same mechanistic steps. For example, the direct interaction between Cas1 and AddAB has not been shown, but it was proposed that Cas1-Cas2 captures the degradation product from AddAB.for spacer integration (Modell et al., 2017).

Among the pull-down candidates, Cas9 showed the highest fold change enrichment (8.8-fold), which is supported by a previous work demonstrating the role of Cas9 in the selection of

prespacers with canonical PAMs in type II-A spacer acquisition (Heler et al., 2015). Cas1 was also highly enriched (4.1-fold) in the pull-down assay, which is presumably due to the Cas1 dimer formation. Surprisingly, Cas2 was not pulled-down by Cas1 bait, although Cas1 and Cas2 are known for forming a complex that is needed for spacer acquisition (Ka et al., 2018; Nunez et al., 2014; Xiao et al., 2017). One possible explanation to this is that the CPD-His_{12}-tag of Cas1 hinders the interaction of Cas1 with Cas2 during the pull-down assay.

The NER protein, UvrB, showed interaction with Cas1 in the pull-down approach with a fold change of 3.5. This finding correlates with the Y2H results, as UvrA was shown to interact with Cas2 and Cas9. Nevertheless, the Cas1-UvrB interaction was not detected in the Y2H assay. This could be attributed to several reasons, such as UvrB possibly having been considered as false negatives due to the stringent condition used for suppressing the auto-activating activity of Cas1 or a non-native testing environment (yeast cellular compartment) in Y2H. Furthermore, the DNA binding domain fused to Cas1 might sterically block the interaction with UvrB. This is the first time that the interactions between type II-A Cas proteins and UvrAB proteins are shown. More studies are needed to investigate the details and the significance of these interactions.

Conforming to an earlier work in the type I-E system (Babu et al., 2011), our pull-down assay demonstrated the interaction of the type II-A Cas1 and AddAB (the Gram-positive paralogs of RecBCD) (3.4-fold and 3.1-fold respectively) from the HR pathway, and MutS (3.7-fold) from the MMR pathway. It was suggested that in both type II-A and type I-E systems, Cas1-Cas2 complex captures the degradation products generated from AddAB (type II-A) or RecBCD (type I-E) for spacer incorporation (Levy et al., 2015; Modell et al., 2017). With regard to this, our result suggests that Cas1 possibly interacts with AddAB during the process of capturing the degradation products from AddAB. As for MutS, the biological significance of its interaction with Cas1 remains to be elucidated.

Altogether, we showed for the first time the physical interactions of Cas1, Cas2 and Cas9 with UvrAB from the NER pathway. More studies are needed to investigate the details and significance of the interactions between the these Cas proteins with UvrAB. Additionally, we also demonstrated the direct interaction of Cas9 and Cas1 for the first time, which was only shown indirectly previously (Heler et al., 2015). Via the pull-down approach we demonstrated

the direct interaction of Cas1 of the type II-A system with AddAB from the HR pathway, which has only been shown in the type I-E system of *E. coli* so far (Babu et al., 2011).

3.3.1.3 The Cas1 SPOT peptide assay identifies the dimerization region of Cas1 and the interacting region of Cas1 with Cas9

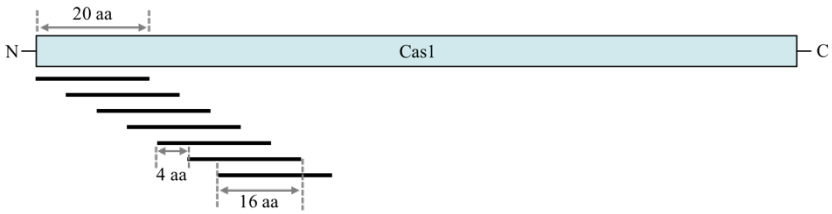
Our Y2H and pull-down assays revealed that the interacting partners of the type II-A Cas proteins are mainly from the CRISPR-Cas systems and the DNA repair pathways. Among the candidates, we were particularly interested in the interaction between Cas9 and Cas1. In the type II-A systems, Cas9 selects prespacers with PAMs that will be used for spacer incorporation by Cas1-Cas2 complex (Heler et al., 2015). We hypothesized that Cas9 interacts with Cas1 possibly via Cas1-Cas2 complex during the prespacers selection process in order to transfer the selected prespacer with PAM to Cas1-Cas2. To address this hypothesis, we used SPOT peptide assay to identify the interacting regions of Cas1 with Cas9, and additionally identify the Cas1 dimerization regions.

SPOT peptide assay is an assay that allows rapid and direct screening of protein-peptide interactions and identification of protein interaction regions. This technique involves incubation of a tagged-interacting protein with a SPOT peptide membrane that contains the peptide arrays of the protein of interest, followed by antibodies incubation and identification of the interacting peptides via the chemiluminescent signals detection. Here, we used a Cas1 SPOT peptide membrane consisting of 72 overlapping peptides, which represent the entire amino acids sequence of Cas1 (**Figure 14A**). On this membrane, each peptide contains 20 amino acids with an overlap of 16 amino acids.

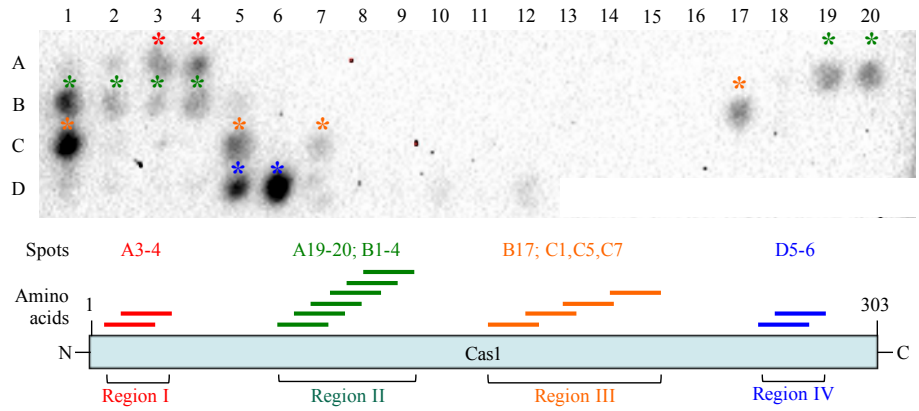
To identify the Cas1-Cas1 dimerization region, a recombinant C-terminal His_6-tagged Cas1 protein was incubated with Cas1 SPOT membrane. Positive signals were detected in four regions that were labeled as I, II, III and IV (**Figure 14B**). However, both of the regions I (residues 9 to 32) and IV (residues 257 to 280) are likely to be false positives, as they contain only two overlapping peptides. Since every peptide (20 amino acids) overlaps with the preceding and succeeding peptides by 16 amino acids, the peptide sequences in the two overlapping regions are also partially present in the continuous overlapping peptides preceding and succeeding the two peptides (**Figure 14A**). Therefore, in principle, the continuous neighboring overlapping peptides should also show positive signals in order to be considered

as a true positive interaction. The same applies to the region III (residues 145 to 204), it was regarded as a false positive due to the fact that the overlapping peptides are non-continuous. The region II (residues 73 to 112; β-sheet-8 to α-helix-3) most likely represents the dimerization site, because this region shows positive signals of a row of 6 continuous overlapping peptides that complies with the mentioned requirement for a true positive (**Figure 14B; Supplementary Figure S4A**).

A



B



C

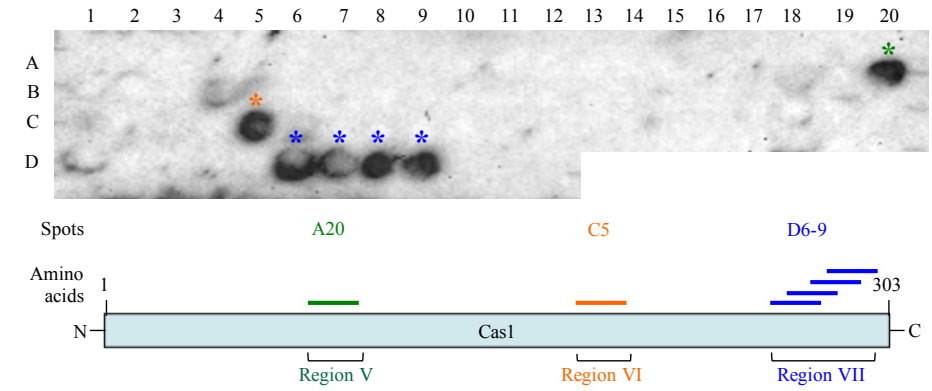


Figure 14. SPOT peptide assay with Cas1.

(A) Scheme showing an array of overlapping peptides (black horizontal lines) of a protein with N- and C-termini indicated. These overlapping peptides are chemically synthesized and immobilized to a cellulose membrane (SPOT peptide membrane) via SPOT peptide synthesis technology. The Cas1 SPOT membrane (synthesized by JPT Peptide Technologies GmbH) consists of 72 overlapping peptides of 20 amino acids. Each peptide overlaps with its preceding peptide by 16 amino acids, *i.e.* each peptide is shifted by 4 amino acids. aa, amino acid. **(B)** The dimerization regions of Cas1, and **(C)** the regions of Cas1 that interact with Cas9. Cas9 was pre-incubated with tracrRNA. **(B-C)** The upper panel shows signals of the Cas1 SPOT membrane after incubating with either Cas1 or Cas9 protein, where the dark spots represent positive signals resulting from the interaction between the Cas1 peptides and the incubated protein. The interacting regions are mapped to Cas1 and indicated at the lower panel. The horizontal numbers 1 to 20 and the vertical letters A to D indicate the coordinate of the spots, where A1 is the first peptide at the N-terminal of Cas1, whereas D12 is the last peptide at the C-terminal. Regions I to VII represent clusters of positive signals.

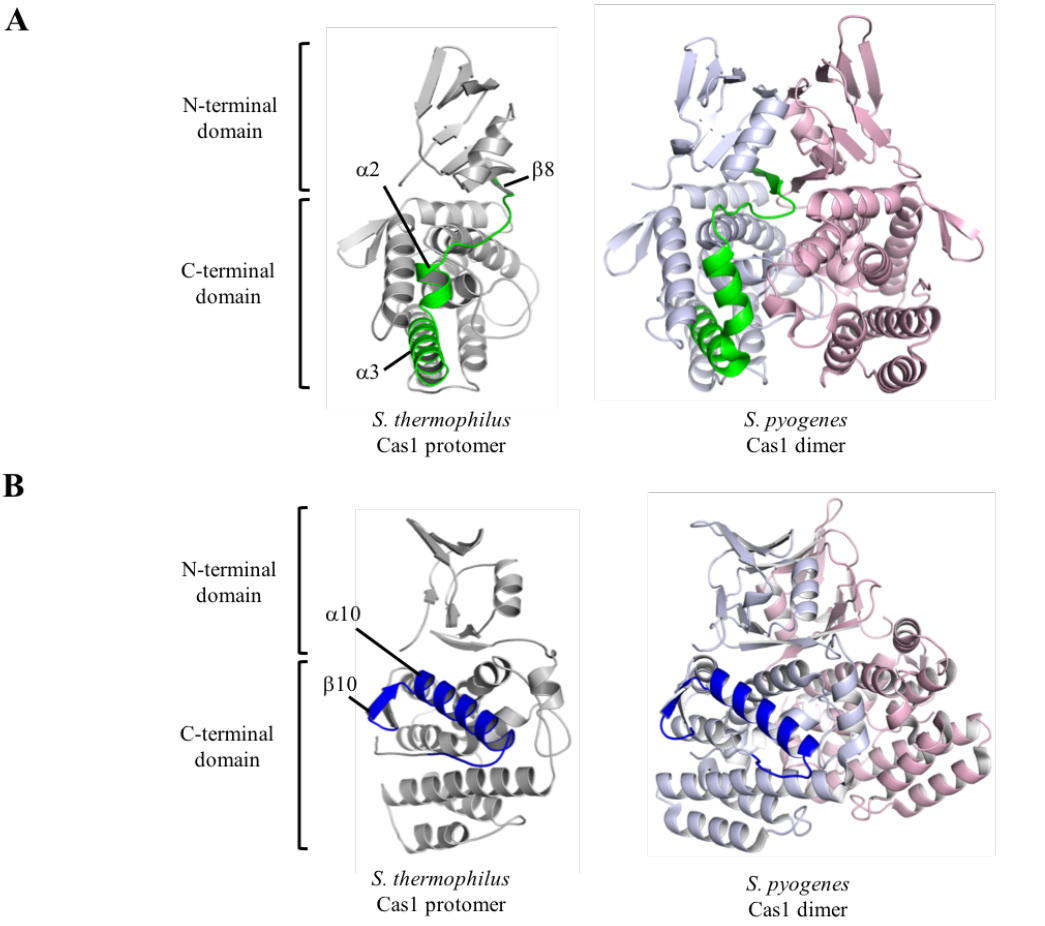
Next, a Cas1 SPOT peptide membrane was tested with a recombinant N-terminal His_6-tagged Cas9 protein pre-incubated with tracrRNA (abbreviated as tracrRNA-Cas9), to pinpoint the interacting region of Cas1 with Cas9. tracrRNA-Cas9 rather than Cas9, was used in this assay, as tracrRNA is critical for spacer acquisition, crRNA biogenesis and interference (Deltcheva et al., 2011; Heler et al., 2015; Jinek et al., 2012). Results show that the regions V (residues 77 to 96) and VI (residues 177 to 196), with a single positive signal each, are likely false positives. The C-terminus of Cas1, region VII (residues 261 to 292; β-sheet-10 to α-helix-10), is most likely the region of Cas1 that interacts with Cas9, as this region demonstrates positive signals of 4 continuous overlapping peptides (**Figure 14C; Supplementary Figure S4B**).

3.3.1.4 Superimposition of the dimerization region of Cas1 and the interacting region of Cas1 with Cas9

While SPOT peptide assay allows rapid identification of the Cas1 dimerization site and Cas1 interacting sites with Cas9, this technique is solely based on the analysis of sequence-dependent interaction, and it has a limitation for structure-dependent interaction. To analyze the interacting regions in the structural perspective, the interacting regions were superimposed on a model structure of *S. thermophilus* LMD-9 Cas1, which was modeled by using *S. pyogenes* Cas1 crystal structure as a template (PDB: 4ZKJ) (Ka et al., 2016) (**Figures 15A-B**). Comparison of the Cas1 dimerization interface of the *S. thermophilus*, β-sheet-8 to α-helix-3 (region II), with

Cas1 dimer of *S. pyogenes* showed that the β-sheet-8 and the loop connecting β-sheet-8 and α-helix-2 matches the dimer interface of *S. pyogenes* Cas1 (**Figure 15A**). To have a closer look on the Cas1 dimerization region the Cas1-Cas2-prespacer complex, we superimposed the interacting region from our results, onto the Cas1-Cas2-prespacer structure of the type II-A system of *E. faecalis* (PDB: 5XVN) (Xiao et al., 2017) (**Figure 15C**). Similar to the Cas1 dimer of *S. pyogenes*, the β-sheet-8 and the connecting loop obtained from the SPOT peptide assay are matching the dimerization region of the Cas1 of *E. faecalis*. Thereby, we provided the evidence that β-sheet-8 and the connecting loop is the dimerization region of the *S. thermophilus* Cas1 by showing that this region matches to the dimerization region of the Cas1 of *S. pyogenes* and *E. faecalis*. As for the interacting region of Cas1 with Cas9 (Cas9 was pre-incubated with tracrRNA), superimposition showed that β-sheet-10 and α-helix-10 (region VII) is located on the outer surface of Cas1 (**Figures 15B-C**), which, in principle, Cas9 could access this region to interact with Cas1. Furthermore, region VII does not overlap with the putative Cas1 dimerization site (**Figure 15A**) and the N-terminal interacting sites of Cas1 with Cas2 (Xiao et al., 2017) or Csn2 (Ka et al., 2016) reported in previous studies. Therefore, it is less possible that the binding of Cas9 to Cas1 would be blocked by the binding of Cas2 and Csn2, respectively, to Cas1.

After confirming the interacting regions via structural comparison, we asked whether the interaction between Cas9 and Cas1 is important for spacer acquisition. Therefore, we investigated the critical residues involved in these interactions via multiple sequence alignment, as this information could be used to further study the impact of the interaction on spacer acquisition. Multiple sequence alignment of Cas1 proteins from several type II-A systems showed that the Cas1 dimerization region includes four strictly conserved residues (P73, Q93, W96 and K112 of Cas1) along with several other highly conserved residues (G82, L90, L94, K101, A104, W105, Q106 and V108) (**Figure 16; green box**). The interacting region of Cas1 with Cas9 is located in the less conserved C-terminus of Cas1, and comprises numerous highly conserved residues (V268, T269, A271, M272, Y275, I280 and I283) (**Figure 16; blue box**). Mutants of these identified residues could be generated via site-directed mutagenesis, verified for their interactions (*i.e.* Cas1 dimerization and Cas9-Cas1 interactions) and subsequently tested for their spacer acquisition activities by using the plasmid-based spacer acquisition assay that was described in the previous section of this thesis.



E. faecalis Cas1-Cas2-prespacer complex

Figure 15. Superimposition of the interacting regions of *S. thermophilus* Cas1 on the *S. pyogenes* Cas1 and *E. faecalis* Cas1-Cas2-prespacer complex.

(A-B) The potential interacting regions are compared to a model structure of *S. thermophilus* LMD-9 Cas1 protomer (CRISPR1) (grey), which was modeled based on the existing crystal structure of Cas1 of *S. pyogenes* (PDB: 4ZKJ) (Ka et al., 2016) (RMSD: 0.76; TM-score 0.896). The *S. pyogene*s Cas1 dimer is shown in white-blue and pink. **(A)** The region II (green; residues 73 to 112) of *S. thermophilus* Cas1 is compared with *S. pyogene*s Cas1 to verify the dimerization regions. The color scheme of the interacting regions is same as in **Figure 14B** and **14C**. **(B)** The interacting region of *S. thermophilus* Cas1 with Cas9, *i.e.* region VII (blue; residues 261 to 292), is superimposed on the *S. pyogene*s Cas1 dimer. **(C)** A structural study demonstrated that the Cas1-Cas2-prespacer complex of *E. faecalis* is formed by the interactions of one Cas2 dimer (light orange and pale yellow) bridging two Cas1 dimers, of which each of the Cas1 dimers is composed of one catalytic Cas1 subunit (pale cyan) and one non-catalytic Cas1 subunit (brown) (PDB: 5XVN) (Xiao et al., 2017). The prespacer is a 22-bp duplex (dark orange) flanked by 4-nt 3′-overhangs and 2-nt 5′-overhangs. The interacting interfaces are only labelled on one of the Cas1 dimers. The interacting region of Cas1 with Cas9 obtained by SPOT peptide assay, is superimposed on Cas1-Cas2-prespacer structure and is labelled as blue and purple on the catalytic and the non-catalytic Cas1 subunit, respectively. The dimerization interface of Cas1 obtained by SPOT peptide assay is labeled as green on the catalytic Cas1 subunit only.

In summary, we identified the dimerization region of Cas1 and interacting region of Cas1 with Cas9 via SPOT peptide assay, and further confirmed the results through structural comparison. By using multiple sequence alignment of type II-A Cas1 proteins, we pinpointed the conserved residues lying in these interacting regions, which provide the basis for studying the biological significance of these residues in Cas9-Cas1 interaction and spacer acquisition.

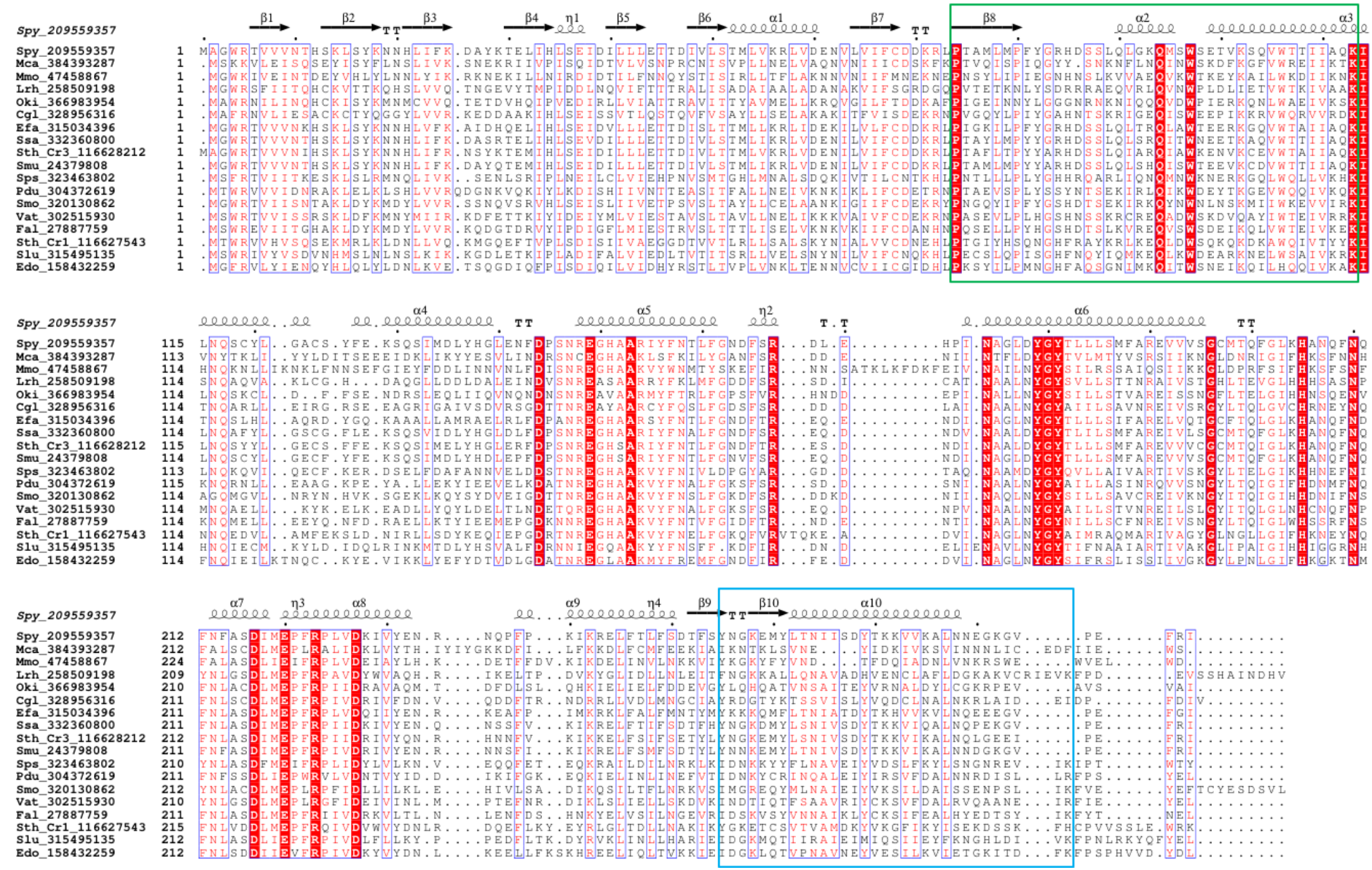

Spy_209559357
Spy_209559357 1 MAGWRTVVVNTHSKLSYKNNHLIFK.DAYKTELIHLSEIDILLLETTDIVLSTMLVKRLVDENVLVIFCDDKRLPTAMLMPFYGRHDSSLQLGKQMSWSETVKSQVWTTIIAQKI
Mca_384393287 1 .MSKKVLEISQSEYISYFLNSLIVK.SNEKRIIVPISQIDTVLVSNPRCNISVPLLNELVAQNVNIIICDSKFKPTVQISPIQGYY.SNKNFLNQINWSKDFKGFVWREIIKTKI
Mmo_47458867 1 .MGWKIVEINTDEYVHLYLNNLYIK.RKNEKILLNIRDIDTILFNNQYSTISIRLLTFLAKNNVNIIFMNEKNEPNSYLIPIEGNHNSLKVVAEQVKWTKEYKAILWKDIIKNKI
Lrh_258509198 1 .MGWRSFIITQHCKVTTKQHSLVVQ.TNGEVYTMPIDDLNQVIFTTTRALISADAIAALADANAKVIFSGRDGQPVTETKNLYSDRRRAEQVRLQVNWPLDLIETVWTKIVAAKI
Oki_366983954 1 .MAWRNILINQHCKISYKMNMCVVQ.TETDVHQIPVEDIRLLVIATTRAVITTYAVMELLKRQVGILFTDDKAEPIGEINNYLGGGNRNKNIQQQVDWPIERKQNLWAEIVKSKI
Cgl_328956316 1 .MAFRNVLIESACKCTYQGGYLVVR.KEDDAAKIHLSEISSVTLQSTQVFVSAYLLSELAKAKITFVISDEKRNPVGQYLPIYGAHNTSKRIGEQISWEEPIKKRVWQRVVRDKI
Efa_315034396 1 .MGWRTVVVNKHSKLSYKNNHLVFK.AIDHQELIHLSEIDVLLLETTDISLSTMLLKRLIDEKILVLFCDDKRLPIGKILPFYGRHDSSLQLTRQLAWTEERKGQVWTAIIAQKI
Ssa_332360800 1 .MGWRTVVVNTHSKLSYKNNHLIFK.DASRTELIHLSEVDILLLETTDIVLSTMLIKRLVDENILVIFCDDKRLPTAYLMPYYGRHDSSLQLSRQITWNEETKAQVWTTIIAQKI
Sth_Cr3_116628212 1 MAGWRTVVVNIHSKLSYKNNHLIFR.NSYKTEMIHLSEIDILLLETTDIVLTTMLVKRLVDENILVIFCDDKRLPTAFLTPYYARHDSSLQIARQIAWKENVKCEVWTAIIAQKI
Smu_24379808 1 .MGWRTVVVNTHSKLSYKNNHLIFK.DAYQTEMIHLSEIDILLLETTDIVLSTMLIKRLVDENILVIFCDDKRLPTAMLMPYYARHDSSLQLSHQISWTEEVKCDVWTTIIAQKI
Sps_323463802 1 .MSFRTVIITKESKLSLRMNQLIVK..SENLSRIPLNEILCLVIEHPNVSMTGHLMNALSDQKIVTILCNTKHLPNTLLLPLYGHHRQARLIQNQMNWKNERKGQLWQLLVKHKI
Pdu_304372619 1 .MTWRVVVIDNRAKLELKLSHLVVRQDGNKVQKIYLKDISHIIVNTTEASITFALLNEIVKNKIKLIFCDETRNPTAEVSPLYSSYNTSEKIRLQIKWDEYTKGEVWQQIVKQKI
Smo_320130862 1 .MGWRTVVIISNTAKLDYKMDYLVVR.SSNQVSRVHLSEISLIIVETPSVSLTAYLLCELAANKIGIIFCDEKRYPNGQYIPFYGSHDTSEKIRKQYNWNLNSKMIIWKEVVIRKI
Vat_302515930 1 .MSWRTVVISSRSKLDFKMNYMIIR.KDFETTKIYIDEIYMLVIESTAVSLTAVLLNELIKKKVAIVFCDEKRNPASEVLPLHGSHNSSKRCREQADWSKDVQAYIWTEIVRRKI
Fal_27887759 1 .MSWREVIITGHAKLDYKMDYLVVR.KQDGTDRVYIPDIGFLMIESTRVSLTTTLISELVKNKVKIIFCDANHNPQSELLPYHGSHDTSLKVREQVSWSDEIKQLVWTEIVKEKI
Sth_Cr1_116627543 1 .MTWRVVHVSQSEKMRLKLDNLLVQ.KMGQEFTVPLSDISIIVAEGGDTVVTLRLLSALSKYNIALVVCDNEHIPTGIYHSQNGHFRAYKRLKEQLDWSQKQKDKAWQIVTYYKI
Slu_315495135 1 .MSWRIVYVSDVNHMSLNLNSLKIK.KGDLETKIPLADIFALVIEDLTVTITSRLLVELSNYNILVIFCNQKHIPECSLQPISGHFNQYIQMKEQLKWDEARKNELWSAIVKRKI
Edo_158432259 1 .MGFRVLYIENQYHLQLYLDNLKVE.TSQGDIQFPISDIQILVIDHYRSTLTVPLVNKLTENNVCVIICGIDHIPKSYILLPMNGHFAQSGNIMKQITWSNEIKQILHQQIVKAKI
Spy_209559357
Spy_209559357 115 LNQSCYL..GACS.YFE.KSQSSIMDLYHGLENFDPSNREGHAARIYFNTLFGNDFSR...DL.E.........HPI.NAGLDYGYTLLLSMFAREVVVSGCMTQFGLKHANQFNQ
Mca_384393287 113 VNYTKLI..YYLDITSEEEIDKLIKYYESVLINDRSNGEGHAAKLSFKILYGANFSR...DD.E.........NII.NTFLDYGYTVLMTYVSRSIIKNGLDNRIGIFHKSFNNH
Mmo_47458867 114 HNQKNLLIKNKLFNNSEFGIEYFDDLINNVNLFDISNREGHAAKVYWNMTYSKEFIR...NN.SATKLKFDKFEIV.NAILNYGYSILRSSAIQSIIKKGLDPRFSIFHKSFSNF
Lrh_258509198 114 SNQAQVA..KLCG.H...DAQGLLDDLDALEINDVSNREASAARRYFKLMFGDDFSR...SD.I.........CAT.NAALNYGYSVLLSTTNRAIVSTGHLTEVGLHHHSASNP
Oki_366983954 114 LNQSKCL..D..F.FSE.NDRSLEQLIIQVNQNDNSNREAVAARMYFTRLFGPSFVR...HNDD.........EPI.NALLNYGYSILLSTVARFISSNGFLTELGIHHNSQENV
Cgl_328956316 114 TNQARLL..EIRG.RSE.EAGRIGAIVSDVRSGDTTNREAYAARCYFQSLFGDSFSR...DD.D.........LAI.NAALNYGYAIILSAVNREIVSRGYLTQLGVCHRNEYNQ
Efa_315034396 114 TNQSLHL..AQRD.YGQ.KAAALLAMRAELRLFDPANREGHAARSYFNTLFGNDFTR...EQ.E.........NDI.NAGLNYGYTLLLSIFAREVQTGCFTQLGLKHANQFND
Ssa_332360800 114 LNQAFYL..GSCG.FLE.KSQSVIDLYHGLDLFDPSNREGHAARIYFNALFGNDFSR...EQ.D.........NDV.NAALDYGYTLILSMFAREIVLSGCMTQFGLKHANQFNQ
Sth_Cr3_116628212 115 LNQSYYL..GECS.FFE.KSQSIMELYHGLERFDPSNREGHSARIYFNTLFGNDFTR...ES.D.........NDI.NAALDYGYTLLLSMFAREVVVCGCMTQIGLKHANQFNQ
Smu_24379808 114 LNQSCYL..GECF.YFE.KSQSIMDLYHDLEPFDPSNREGHSARIYFNTLFGNVFSR...EQ.D.........NDI.NAGLDYGYTLLLSMFAREVVVSGCMTQFGLKHANQFNQ
Sps_323463802 113 LNQKQVI..QECF.KER.DSELFDAFANNVELDDSTNREGHAAKVYFNIVLDPGYAR...GD.D.........TAQ.NAAMDYGYQVLLAIVARTIVSKGYLTELGIKHHNEFNI
Pdu_304372619 115 KNQRNLL..EAAG.KPE.YA.LLEKYIEEVELKDATNREGHAAKVYFNALFGKSFSR...SD.D.........SNI.NAALNYGYAILLASINRQVVSNGYLTQIGIFHDNMFNQ
Smo_320130862 114 AGQMGVL..NRYN.HVK.SGEKLKQYSYDVEIGDTTNREGHAAKVYFNSLFGKDFSR...DDKD.........NII.NAQLNYGYSILLSAVCREIVKNGYITQIGIHHDNIFNS
Vat_302515930 114 MNQAELL...KYK.ELK.EADLLYQYLDELTLNDETQREGHAAKVYFNALFGKSFSR...EQ.D.........NPV.NAALNYGYAILLSTVNREILSLGYITQLGLNHCNQFNP
Fal_27887759 114 KNQMELL..EEYQ.NFD.RAELLKTYIEEMEPGDKNNREGHAAKVYFNTVFGIDFTR...ND.E.........NTI.NAALNYGYNILLSCFNREIVSNGYLTQIGLWHSSRFNS
Sth_Cr1_116627543 114 NNQEDVL..AMFEKSLD.NIRLLSDYKEQIEPGDRTNREGHAAKVYFNELFGKQFVRVTQKE.A.........DVI.NAGLNYGYAIMRAQMARIVAGYGLNGLLGIFHKNEYNQ
Slu_315495135 114 HNQIECM..KYLD.IDQLRINKMTDLYHSVALFDRNNIEGQAAKYYFNSFF.KDFIR...DN.D.........ELIENAVLNYGYTIFNAAIARTIVAKGLIPAIGIHHIGGRNH
Edo_158432259 114 FNQIEILKTNQC..KYE.VIKKLYEFYDTVDLGDATNREGLAAKMYFREMFGNDFIR...FE.D.........DVI.NAGLNYGYSIFRSLISSIIVGKGYLPNLGIFHKGKTNM
Spy_209559357
Spy_209559357 212 FNFASDIMEPFFRPLVDKIVYEN.R....NQPFP...KIKRELFTLFSDTFSYNGKEMYLTNIISDYTKKVVKALNNEGKGV.....PE.....FRI.........
Mca_384393287 212 FALSCDLMEPLRALIDKLVYTH.IYIYGKKDFI...LFKKDLFCMFEEKIAIKNTKLSVNE...YIDKIVKSVINNNLIC..EDFIIE.....WS.........
Mmo_47458867 224 FALASDLIEIFRPLVDEIAYLH.K....DETFFDV.KIKDELINVLNKKVIYGKYFYVND....TFDQIADNLVNKRSWE.....WVEL.....WD.........
Lrh_258509198 209 YNLGSDLMEPFRPAVDYWVAQH.R....IKELTP..DVKYGLIDLLNLEITFNGKKALLQNAVADHVENCLAFLDGKAKVCRIEVKFPD.....EVSSHAINDHV
Oki_366983954 210 FNLACDLMEPFRPIIDRAVAQM.T....DFDLSL..QHKIELIELFDDEVGYLQHQATVNSAITEYVRNALDYLCGKRPEV.....AVS.....VAI.........
Cgl_328956316 211 FNLSCDLMEPFRPLIDFDVN.V....QDDFTR...NDRRLLVDLMNGCIAYRDGTYKTSSVISLYVQDCLNALNKRLAID...EIDP.....FDIV.........
Efa_315034396 211 FNLASDLMEPFRPLVDQIIYEN.R....KEAFP...IMKRKLFALFMNTYMYKKKQMFLTNIATDYTKHVVKVLNQEEEGV.....PE.....FGI.........
Ssa_332360800 211 FNLASDIMEPFRPIIDKIVYEN.Q....NSSFV...KIKRELFTIFSDTFHYNGKDMYLSNIVSDYTKKVIQALNQPEKGV.....PE.....FRI.........
Sth_Cr3_116628212 212 FNLASDIMEPFRPIIDRIVYQN.R....HNNFV...KIKKELFSIFSETYLYNGKEMYLSNIVSDYTKKVIKALNQLGEEI.....PE.....FRI.........
Smu_24379808 211 FNFASDIMEPFRPIVDRIVYEN.R....NNSFI...KIKRELFSMFSDTYLYNNKEMYLTNIVSDYTKKVIKALNNDGKGV.....PE.....FRI.........
Sps_323463802 210 YNLASDFMEIFRPLIDYLVLKN.V....EQQFET..EQKRAILDILNRKLKIDNKKYYFLNAVEIYVDSLFKYLSNGNREVV...IKIPT.....WTY.........
Pdu_304372619 211 FNFSSDLIEPWRVLVDNTVYID.D....IKIFGK..EQKIELINLINEFVTIDNKYCRINQALEIYIRSVFDALNNRDISL...LRFPS.....YEL.........
Smo_320130862 212 YNLACDLMEPLRPFIDLLILKL.E....HIVLSA..DIKQSILTFLNRKVSIMGREQYMLNAIEIYVKSILDAISSENPSL...IKFVE.....YEFTCYESDSVL
Vat_302515930 210 YNLGSDLMEPLRGFIDEIVINL.M....PTEFNR..DIKLSLIELLSKDVKINDTIQTFSAAVRIYCKSVFDALRVQAANE...IRFIE.....YEL.........
Fal_27887759 211 FNLASDLMEPFRRIVDRKVLTL.N....LENFDS..HNKYELVSILNGEVRIDSKVSYVNNAIKLYCKSIFEALHYEDTSY...IKFYT.....NEL.........
Sth_Cr1_116627543 215 FNLVDDLMEPFRRQIVDVWVYDNLR....DQEFLKY.EYRLGLTDLLNAKIKYGKETCSVTVAMDKYVKGFIKYISEKDSSK...FHCPVVSSLE.WRK.........
Slu_315495135 212 FNLASDLIEPFRPIVDLFLLKY.P....PEDFLTK.DYRVKLINLLHARIEIDGKMQTIIRAIEIMIQSIIEYFKNGHLDI...VKFPNLRKYQFYEL.........
Edo_158432259 212 FNLSDDIIEVFRPIVDKYVYDN.L....KEELLFKSKHREELIQLTVKKIEIDGKLQTVPNAVNEYVESLLKVLETGKITD...FKFPSPHVVD.YDL.........

Figure 16. Multiple sequence alignment of the Cas1 proteins of the type II-A CRISPR-Cas systems.

Green box (residues 73 to 112 of Sth_Cr1; β-sheet-8 to α-helix-3; region II) highlights the dimerization region of Cas1 based on the results of SPOT assay, whereas blue box (residues 261 to 292 of Sth_Cr1; β-sheet-10 to α-helix-10; region VII) indicates the interacting region of Cas1 with Cas9 that was pre-incubated with tracrRNA. Cas1 sequences are labelled by their species names and GenInfo (GI) identifier. Mca, *Mycoplasma canis* PG 14; Mmo, *Mycoplasma mobile* 163K; Lrh, *Lactobacillus rhamnosus* GG; Oki, *Oenococcus kitaharae* DSM 17330; Cgi, *Coriobacterium glomerans* PW2; Ssa, *Streptococcus sanguinis* SK49; Sth_Cr3, *Streptococcus thermophilus* LMD-9, CRISPR3; Spy, *Streptococcu pyogenes* SF370 (M1 GAS); Smu, *Streptococcus mutans* UA159; Sps, *Staphylococcus pseudintermedius* ED99; Pdu, *Peptoniphilus duerdenii* ATCC BAA-1640; Smo, *Solobacterium moorei* F0204; Vat, *Veillonella atypica* ACS-134-V-Col7a; Fal, *Filifactor alocis* ATCC 35896; Sth_Cr1, *Streptococcus thermo*philus LMD-9, CRISPR1; Efa, *Enterococcus faecalis* TX0027; Slu, *Staphylococcus lugdunensis* M23590; Edo, *Eubacterium dolichum* DSM 3991. The multiple sequence alignment was created using MUSCLE (3.8) (Edgar, 2004a, 2004b) and ESPRIPT (Robert and Gouet, 2014). Red box with white characters, strictly conserved residues; Blue frame with red characters, highly conserved residues.

3.3.2 Investigation of Cas9-Cas1 interaction

3.3.2.1 Interaction studies of Cas9, Cas1 and Cas2 via size-exclusion chromatography

Based on a previous study, the interactions between Cas9, Cas1, Cas2 and Csn2 were known (Heler et al., 2015), yet it was unclear whether Cas9 directly interacts with Cas1. Here, our study revealed the direct interaction of Cas9 and Cas1. However, it is still unclear how Cas9 interacts with Cas1. Structural studies reported a stable Cas1-Cas2 complex formation in the type I-E and type II-A systems (Nunez et al., 2014; Xiao et al., 2017) and Cas1–Cas2-Cas3 (Cas2 is naturally fused to Cas3) complex formation in the type I-F system (Fagerlund et al., 2017). Interestingly, a structural study in the type I-C system of *B. halodurans* showed a direct interaction of Cas4 with Cas1, rather than Cas4 with Cas1-Cas2 complex (Lee et al., 2018). Nevertheless, how does the Cas4-Cas1 complex transform to the final adaptation complex is yet to be elucidated. Hence, we wondered whether the type II-A Cas9 interacts with Cas1 directly or via the Cas1-Cas2 complex. To address this question, SEC was used as one of the approaches to analyze the protein complex formation. Additionally, this technique could be used to establish a read-out system to verify whether the site-directed mutations on the interacting residues obtained from our studies (**Figures 14-16**), could interfere the Cas9-Cas1 interaction. Formation of a complex with higher molecular weight can be identified by comparing the elution profiles of the pre-incubated interacting proteins to the protein profiles of every single individual protein component.

In these SEC studies, the complex formations involved different pre-incubation combinations of N-terminal His_6-tagged Cas9, C-terminal His_6-tagged Cas1 and untagged Cas2 (the N-terminal SUMO-His_6-tag of Cas2 was cleaved during protein purification), which are abbreviated as Cas9, Cas1 and Cas2 here. In order to study whether Cas9 interacts with Cas1 directly or via the Cas1-Cas2 complex, two protein combinations were analyzed via SEC, *i.e.* (1) Cas9, Cas1 and Cas2 and (2) Cas9 and Cas1.

Cas9, Cas1 and Cas2 were individually expressed in *E. coli* because their expression conditions were different (see Materials and Methods). Afterward, the cell lysates of the three proteins were combined prior to the protein purification using immobilized metal affinity chromatography (IMAC) with TALON Co^{2+} resin and SEC. The amount of eluted proteins was

very low as indicated by the elution peaks of the chromatogram that were lower than 13 mAU (mili-Absorbance Units) at UV 280 nm (**Supplementary Figure S5**). SDS-PAGE gel of the eluted fractions demonstrated that Cas2 monomer (12.6 kDa) was eluted as peak 4, whereas Cas1 monomer (36.2 kDa) was eluted as peaks 1, 2 and 3. Based on the observation of our routine protein purifications with SEC, peak 5 was the elution of the imidazole used in the buffers for protein purification with IMAC. Three additional protein bands were detected at peak 3, *i.e.* bands 2, 3 and 4, which might be either co-purified unspecific proteins or Cas2 aggregates. Band 3 could also be a Cas2 dimer that was not totally denatured in the SDS-PAGE gel, as the estimated molecular weight of band 3 is about the same as Cas2 dimer. Peak 1 contained both Cas9 (130.4 kDa) and Cas1. There was neither evidence for co-purification of Cas9, Cas1 and Cas2, nor for co-purification of Cas2 and Cas1. The molecular weight of the additional band, band 1, indicates that band 1 could be the incompletely denatured Cas1 dimer in SDS-PAGE gel. It was unclear whether the co-elution of Cas9 and Cas1 as peak 1 was the outcome of Cas9-Cas1interaction or the overlapping of the elution fractions of Cas1 from the tail of peak 2 with Cas9 from peak 1.

The resolution of the chromatogram for the study of Cas9-Cas1-Cas2 complex formation was not optimal, as the broad peak widths resulted the overlapping of one peak with another (**Supplementary Figure S5**). Therefore, the elution fractions at the borders of the peaks always consisted of more than one proteins. To improve the resolution of SEC, a lower sample load, *i.e.* 200 µl instead of 500 µl, was used in the following SEC (the SEC for Cas9 and Cas1 complex formation). Moreover, a smaller elution volume per fraction, *i.e.* 250 µl instead of 500 µl, was also applied in the following SEC in order to visualize well-defined changes of the protein components along the elution profile in the SDS-PAGE gel.

As for the SEC of the Cas9 and Cas1 complex formation, both proteins were individually expressed and purified via IMAC with TALON Co^{2+} resin and SEC. This modified protocol allowed the flexible adjustment of the molar ratio of the pre-incubated Cas9 and Cas1 to obtain the optimal condition for the interaction. Purified Cas9 and Cas1 were pre-incubated in 1:3 molar ratio before SEC. Here, higher amount of Cas1 was used as Cas1 usually forms a dimer. Analysis on a SDS-PAGE gel showed that Cas1 was eluted as peak 2 and 3, and both Cas9 and Cas1 co-eluted as peak 2 (**Figure 17A**). Mass-spectrometry (data not shown) identified that bands 1 and 2 were Cas9, which were probably Cas9 aggregates (**Figure 17A**). To verify whether the co-eluted Cas9 and Cas1 proteins was a complex, equal amount of Cas1

was analyzed under the same conditions, and the elution profile of Cas1 was compared to the elution profile of the pre-incubated Cas9 and Cas1. The Cas1 only SEC analysis showed that a low amount of Cas1 was also detected in the elution fractions of peak 2 (**Figure 17B**), which was probably due to different oligomerization states of Cas1. In fact, an extra peak that was eluted prior to Cas1 dimer was always observed in the routine SEC purifications of Cas1. The comparison of both elution profiles showed that peak 3 corresponded to the Cas1 elution, whereas peak 2 corresponded to the Cas9 elution together with a lower amount of Cas1 (**Figure 17C**). Therefore, there is no strong evidence showing the interaction of Cas9 and Cas1 in this study.

Altogether, there was no solid observation of protein-protein interaction for the mentioned protein combinations under the tested conditions using SEC. The SEC resolution of the Cas9 and Cas1 combination (**Figures 17B-C**) was improved compared to the SEC resolution of Cas9, Cas1 and Cas2 combination (**Supplementary Figure S5**). Varying other parameters such as the salt concentration and pH of the mobile phase of SEC could optimize the conditions for protein complex formation. Since tracrRNA is essential in spacer acquisition (Heler et al., 2015), pre-incubating Cas9 with tracrRNA prior to the incubation with Cas1 and Cas2 or Cas1 only might facilitate the complex formation. Furthermore, introducing a prespacer with a PAM during the pre-incubation of the proteins mixture (*i.e.* tracrRNA, Cas9, Cas1, with Cas2 in one case and without Cas2 in the other) may also promote the complex formation. The rationale behind this is as follows: a prespacer with a PAM can be recognized by Cas9 in the presence of tracrRNA (Heler et al, 2015), and the PAM recognition could subsequently facilitate the interaction between Cas9 and Cas1 to allow the transfer of the prespacer to the Cas1-Cas2 complex for spacer integration. In addition to SDS-PAGE gel, native polyacrylamide gel could be used in parallel for the direct detection of the protein complexes in their native forms at higher molecular weights. SEC is commonly used for protein purification and also protein complex formation. While this technique is more suitable for stronger protein interaction, other technique such as crosslinking is also commonly used, especially for weak and transient protein interactions.

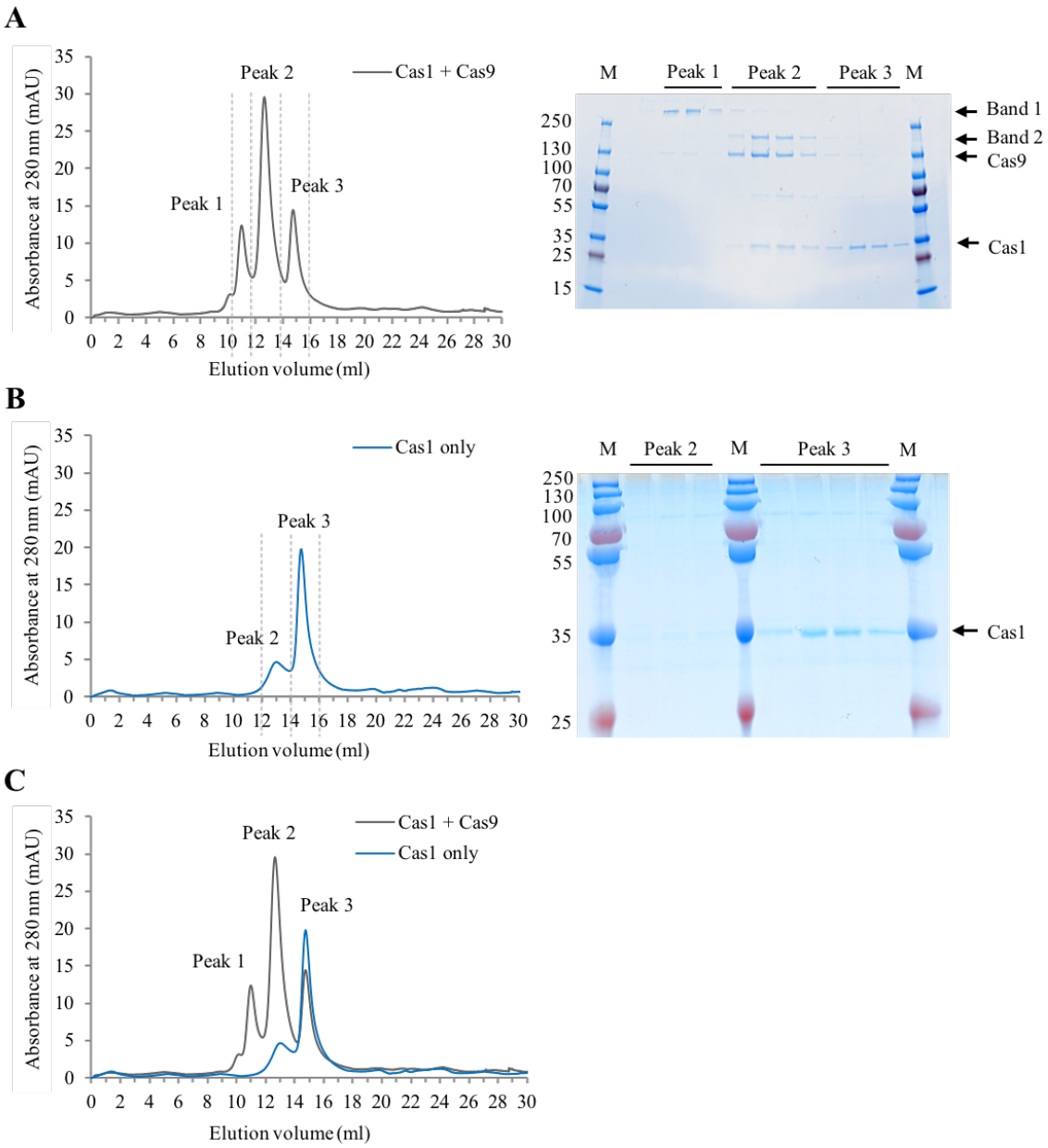


Figure 17. Investigation of Cas9 and Cas1 interaction via size-exclusion chromatography.

The recombinant N-terminal His_6-tagged Cas9 and C-terminal His_6-tagged Cas1 proteins were individually expressed and purified via IMAC with TALON Co^{2+} resin and size-exclusion chromatography (SEC) with the HiLoad® 16/600 Superdex® 200 pg (GE Healthcare). **(A)** After first round of SEC purification, Cas9 (25 µM) and Cas1 (75 µM) were pre-incubated in 1:3 molar ratio before applying to the SEC column Superdex® 200 Increase 10/300 GL (GE Healthcare). Protein eluates obtained by SEC were analyzed by 4-20% Coomassie stained-SDS-PAGE gels. Cas9 elutes as peak 2, whereas Cas1 eluates as peak 2 and 3. Peak 1 is possibly the elution of protein aggregates. **(B)** After first round of SEC purification, an individual Cas1 (75 µM) was analyzed with SEC column Superdex® 200 Increase 10/300 GL (GE Healthcare) using the same conditions as for **(A)**. The eluted fractions were separated by 10% Coomassie stained-SDS-PAGE gels. Cas1 eluates as peak 2 and 3. **(C)** The elution profiles of **(A)** and **(B)** were compared. Cas9, 130.4 kDa; Cas1, 36.2 kDa. M, PageRuler™ Plus Pre-stained Protein Ladder (Thermo Scientific).

3.3.2.2 Crosslinking studies of Cas9, Cas1 and Cas2

In addition to SEC, we used crosslinking as a second approach to address the question regarding whether Cas1-Cas2 complex is needed for Cas9-Cas1 interaction, as well as to establish a read-out system for verifying the interacting residues. Furthermore, we applied crosslinking to confirm the oligomerization states of Cas1 and Cas2. Crosslinking reagents are used to preserve the protein-protein complexes by covalently linking the specific amino acid functional groups on the interacting proteins together upon their interaction. Crosslinking is commonly used to study proteins interactions, especially for weak or transient interactions, and it is also used to assess the distance of the interacting residues (Singh et al., 2010). Bis(sulfosusccinimidy) suberate (BS^3), a commonly used crosslinker, was applied in this study. BS^3 allows the amine-to-amine crosslinking between the interacting proteins via its homobifunctional *N*-hydroxysulfosuccinimide (NHS) ester with an 8-carbon spacer arm. Numerous primary amines usually present in the side chain of lysine residues of a protein and N-terminus of each polypeptide, therefore feasible for NHS-ester crosslinking. The recombinant His-tag Cas proteins used in this crosslinking study were same as those proteins used in SEC.

In order to verify the oligomerization state of individual Cas1 and Cas2 proteins, Cas1 and Cas2 were separately titrated with the BS^3 crosslinker, and the reactions were visualized on Coomassie-stained gels (**Figures 18A-B**). When increasing amount of BS^3 crosslinker was added, both Cas1 and Cas2 showed dimer formation, which confirmed that the Cas1 and Cas2 of the type II-A system (CRISPR1) of S. *thermophilus* LMD-9 form dimers, respectively. These findings are agreement with the previous studies (Babu et al., 2011; Beloglazova et al., 2008; Kim et al., 2013; Nam et al., 2012; Samai et al., 2010; Wiedenheft et al., 2009; Xiao et al., 2017). Due to the innate limited stability of Cas1, a small portion of Cas1 precipitated during the crosslinking reaction. The Cas1 precipitates appeared as bands above 70 kDa.

To investigate the interactions between the different Cas proteins, various combinations (*i.e.* Cas1 and Cas2; tracrRNA-Cas9 and Cas1; tracrRNA-Cas9, Cas1 and Cas2) were pre-incubated before crosslinking. These reactions were divided into three portions and analyzed with Coomassie-stained gradient gel (**Figure 18C**) and Western Blots with Cas1 (**Figure 18D**) and Cas2 antibodies in parallel (**Figure 18E**).

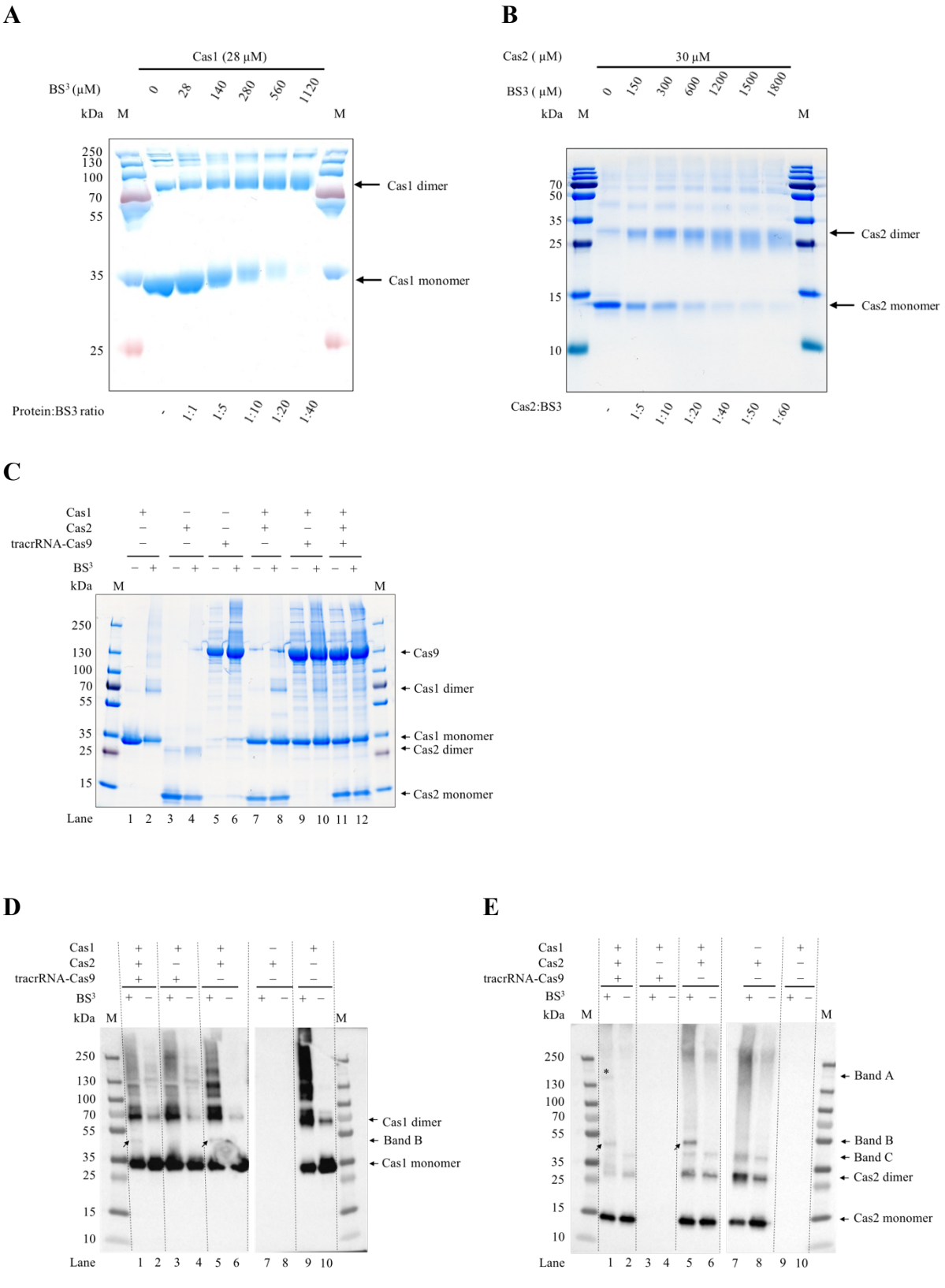

A
Cas1 (28 μM)
BS³ (μM)
0 28 140 280 560 1120
kDa M M
250 130 100 70 55 35 25
Cas1 dimer
Cas1 monomer
Protein:BS3 ratio
- 1:1 1:5 1:10 1:20 1:40
B
Cas2 (μM)
30 μM
BS3 (μM)
0 150 300 600 1200 1500 1800
kDa M M
70 50 35 25 15 10
Cas2 dimer
Cas2 monomer
Cas2:BS3
- 1:5 1:10 1:20 1:40 1:50 1:60
C
Cas1
Cas2
tracrRNA-Cas9
BS³
kDa M M
250 130 100 70 55 35 25 15
Cas9
Cas1 dimer
Cas1 monomer
Cas2 dimer
Cas2 monomer
Lane 1 2 3 4 5 6 7 8 9 10 11 12
D
Cas1
Cas2
tracrRNA-Cas9
BS³
kDa M M
250 130 100 70 55 35 25 15 10
Cas1 dimer
Band B
Cas1 monomer
Lane 1 2 3 4 5 6 7 8 9 10
E
Cas1
Cas2
tracrRNA-Cas9
BS³
kDa M M
250 130 100 70 55 35 25 15 10
Band A
Band B
Band C
Cas2 dimer
Cas2 monomer
Lane 1 2 3 4 5 6 7 8 9 10

Figure 18. The study of protein-protein interaction via crosslinking assays.
(A) Cas1-Cas1 crosslinking. 28 μM Cas1 was titrated with the BS^3 crosslinker. The crosslinked reaction was analyzed on 10% Coomassie stained-SDS-PAGE gels. **(B)** Cas2-Cas2 crosslinking. 30 μM Cas2 was titrated with the BS^3 crosslinker. The crosslinked reaction was analyzed on 15% Coomassie stained-SDS-PAGE gels. **(C)** Three different protein combinations were used for BS^3 crosslinking study, *i.e.* (1) Cas1 and Cas2, (2) tracrRNA-Cas9 and Cas1, and (3) tracrRNA-Cas9, Cas1 and Cas2. tracrRNA-Cas9 was prepared by pre-incubating these two components in 2:1 molar ratio at 37°C for 10 min before the incubation with other proteins. The individual protein components (20 μM each) were mixed pre-incubated for one hour at 4°C before the addition of BS^3 crosslinker. The molar ratio of individual proteins is 1:1, whereas the molar ratio of individual proteins to BS^3 is 1:4, *i.e.* 20 μM individual proteins and 80 μM of BS^3 crosslinker were used. The crosslinked proteins mixture was analyzed on the 4-20% Coomassie stained-SDS-PAGE gels. **(D)** The crosslinking reactions in **(C)** were analyzed on Western Blots with Cas1 antibody. **(E)** The crosslinking reactions in **(C)** were analyzed on Western Blots with Cas2 antibody. Cas9 monomer, 130.4 kDa; Cas1 monomer, 36.2 kDa; Cas1 dimer, 72.4 kDa; Cas2 monomer, 12.6 kDa; Cas2 dimer, 25.2 kDa. Estimated sizes for speculated complexes are 170 kDa for $Cas1_4$-$Cas2_2$, 202.8 kDa for Cas9-$Cas1_2$, 300.4 kDa for Cas9-$Cas1_4$-$Cas2_2$. Asterisk, band A; black arrow, band B; M, PageRuler™ Plus Pre-stained Protein Ladder (Thermo Scientific).

A likely intermediate interaction product of the Cas1-Cas2 complex (band B; marked with an arrow) with a molecular weight between 35-55 kDa was observed in the crosslinking reactions that involved both Cas1 and Cas2. (**Figures 18D-E, lanes 1 and 5**). The reported stoichiometry of the Cas1-Cas2 complex is $Cas1_4$-$Cas2_2$ in type II-A system of *E. faecalis* and type I-E system of *E. coli* (Nunez et al., 2014; Xiao et al., 2017), wherein one Cas2 dimer is sandwiched between two Cas1 dimers. Since Cas1 and Cas2 are very conserved in CRISPR-Cas systems (Makarova et al., 2011), the stoichiometry of the Cas1-Cas2 complex in type II-A systems of *S. thermophilus* is assumed to be same as the reported stoichiometry (Nunez et al., 2014; Xiao et al., 2017). This indicate that the molecular weight of the $Cas1_4$-$Cas2_2$ complex of *S. thermophilus* is possibly around 170 kDa. Nevertheless, the band of this molecular weight was not detected by Western Blot in the crosslinking reaction of Cas1 and Cas2 (**Figures 18D-E, lane 5**).

A faint band (band A; marked with an asterisk) with a molecular weight between 130-250 kDa was detected in the Western Blot for the crosslinking reaction of the tracrRNA-Cas9, Cas1 and Cas2 with the Cas2 antibody (**Figure 18E, lane 1**), but not with the Cas1 antibody (**Figure 18D, lane 1**). The molecular weight of the Cas1-Cas2 complex matches the molecular weight of band A. However, it is unclear whether band A is the Cas1-Cas2 complex, as band A was not detected in the crosslinking reaction of Cas1 and Cas2 (**Figure 18D-E, lane 5**). For the

Western Blot with Cas1 antibody, band A might have been masked by the high amount of smear (**Figures 18D-E, lane 5**). The molecular weight of Cas1 is about three times larger than Cas2, which means that Cas1 has more primary amines available for crosslinking reactions. Hence, crosslinking reactions occur more frequently in Cas1 compared to Cas2. The number of crosslinking events in Cas1 could affect the migration speed of the protein in the SDS-PAGE gel, therefore, a higher amount of smearing is visible on the Western Blot analysis with Cas1 antibody (**Figure 18D**). To reduce the smearing background, the amount of the proteins used in the crosslinking reactions could be further reduced and the crosslinking reactions could be directly analyzed with the Western Blot.

Routine Cas2 purification and the subsequent SDS-PAGE gel analysis always showed two faint bands, which were slightly above and below the major Cas2 band. Mass-spectrometry analysis identified that the faint bands were Cas2 (data not shown). These two Cas2 moieties are probably improperly folded Cas2 or degraded Cas2 migrating at different speed. The band C could be the outcome of the crosslinking of these Cas2 moieties (**Figure 18E, lanes 1-2 and 5-8**).

It is not impossible that both bands A (**Figure 18E, lane 1**) and B (**Figures 18D-E, lanes 1 and 5**) are the intermediate crosslinked products of tracrRNA-Cas9-Cas1-Cas2 and Cas1-Cas2, respectively. However, in order to detect the final crosslinked-complexes, if any, it requires the interactions and crosslinking that involves two crosslinked Cas1 dimers, one crosslinked Cas2 dimers and/or one tracrRNA-Cas9. In addition, the amount of the final crosslinked complexes must be above the detection limit of the Western Blot. All these factors add complexity to the detection of the final crosslinked-complex with the crosslinking technique. The addition of protospacer with canonical PAM into the proteins mixture might also facilitate the interactions between tracrRNA-Cas9, Cas1 and Cas2, as per the explanation earlier.

While the *in vitro* crosslinking assay has its limitation in crosslinking multiple proteins in different oligomerization states, other techniques could also be used for studying the interactions of these proteins. For instance, surface plasmon resonance (SPR) is a technology that based on surface plasmon resonance and it is used for studying PPIs. SPR detects the protein interactions via the change of refractive index when a prey is probed on a bait that is immobilized on a sensor chip. The proteins eluted from SPR are further analyzed by mass-spectrometry. However, SPR has its limitation in detecting transient protein interactions

(Brückner et al., 2009). *In vivo* methods such as fluorescence resonance energy transfer (FRET), bioluminescence resonance energy transfer (BRET) and bimolecular fluorescence complementation (BiFC) have the advantage of allowing the detection of proteins interactions in living cells via confocal microscopy or scanning force microscopy (Brückner et al., 2009). Many conditions could be investigated with *in vivo* detection of proteins interactions, for example, the protein interactions of Cas9 with Cas1 with and without phage infection. However, these *in vivo* methods are comparatively laborious.

In summary, the dimer formations of Cas1 and Cas2, respectively, were confirmed via the crosslinking assays, and they are in line with other studies (Babu et al., 2011; Beloglazova et al., 2008; Kim et al., 2013; Nam et al., 2012; Samai et al., 2010; Wiedenheft et al., 2009; Xiao et al., 2017). There is a lack of solid conclusion regarding the complex formation of Cas1-Cas2, tracrRNA-Cas9-Cas1 and tracrRNA-Cas9-Cas1-Cas2.

4 Discussion

4.1 Unravelling type II-A spacer acquisition

In this work, we investigated the essential elements in type II-A spacer acquisition via two different approaches, *i.e.* a heterologous system expressed in the *E. coli* BL21-AI host and an endogenous system. In the absence of Cas9, spacer acquisition of the *S. pyogenes* and *S. thermophilus* heterologous type II-A systems were not detected. This could be explained by the requirement of Cas9 and tracrRNA in spacer acquisition in the type II-A systems which had been published in the course of this study (Heler et al., 2015; Wei et al., 2015a).

Although all Cas proteins and tracrRNA are included in two experimental set-ups for the *S. pyogenes* heterologous type II-A system, we also did not observe spacer acquisition (**Figures 7C and 7D; Supplementary Table S1**). This could be due to the potential limitation of *E. coli* as a host for the heterologous systems for the type II-A spacer acquisition study. Thus far, *E. coli* has only been used as a host for the characterization of spacer acquisition mechanisms of the type I systems originating from different Gram-negative bacteria (Kieper et al., 2018; Vorontsova et al., 2015; Yosef et al., 2012). The type II-A systems are mainly encoded in Firmicutes (primarily Gram-positive bacteria) and they are absent in Proteobacteria (mainly Gram-negative bacteria) (Bernheim et al., 2017). Therefore, it is unclear whether the Gram-negative host, *E. coli,* could provide the native conditions required for spacer acquisition of the type II-A systems.

While our studies were in process, spacer acquisition was shown in the heterologous type II-A systems of *S. pyogenes* SF370 in the Gram-positive host, *S. aureus* RN4420, which also does not encode any CRISPR-Cas system (Heler et al., 2015). This led to the question whether there are any non-CRISPR associated elements encoded in the Gram-positive host required for the type II-A spacer acquisition, which are absent in the Gram-negative host, *E. coli* BL21-AI. For example, AddAB, the homologous recombination proteins that are mainly encoded in the Gram-positive bacteria, was suggested to play a role in the prespacer generation in the type II-A system (Modell et al., 2017). Nonetheless, it is not known whether the interaction between AddAB and Cas1-Cas2 are required to facilitate the prespacer substrates generation and

capturing processes. Our work and a former study revealed the interactions between AddAB and Cas1 of the type II-A systems of *S. thermophilus* (**Figure 13B; Table 2**), and RecBC subunits (subunits of RecBCD complex) and Cas1 of the type I-E system of *E. coli* (Babu et al., 2011), respectively. These findings suggest that the interaction between the homologous recombination proteins and Cas1 (presumably via Cas1-Cas2 complex) is possibly necessary during the prespacer substrates generation and capturing steps. While AddAB complex possesses one helicase and two nuclease domains, RecBCD complex harbors two helicase and one nuclease domains (Wigley, 2013). Taking into the consideration of the structural differences between AddAB and RecBCD, it is possible that the type II-A Cas1-Cas2 complex could not interact with RecBCD. Assuming the above is correct, *E. coli* might not be a suitable host for spacer acquisition study of the type II-A systems that have been hitherto reported in Gram-positive bacteria only. To investigate this hypothesis, future experiment shall focus on determining whether the prespacer substrates generation and capturing processes require the interaction between the type II-A Cas1 and AddAB, and whether the type II-A Cas1 could interacts with RecBCD.

To investigate spacer acquisition under native conditions, the acquisition of new spacers was investigated in *S. pyogenes* using a plasmid challenge approach. However, spacer acquisition was not observed in *S. pyogenes* WT, the Cas9 RuvC mutant (pEC85ΩtracrRNA-*cas9-D10A*-speM) and the Cas9 HNH mutant (pEC85ΩtracrRNA-*cas9-H840A*-speM) (**Figure 8; Supplementary Table S1**). The undetectable spacer acquisition in *S. pyogenes* could be due to several reasons. For instance, it could be associated with the low endogenous expression of the type II-A *cas* genes demonstrated by our RNA sequencing data (data not shown). In this case, it is possible to boost the spacer acquisition activities in *S. pyogenes* by increasing the expression levels of all the Cas proteins. In agreement with that, spacer acquisition activity was increased and detectable in the endogenous type II-A system of *S. thermophilus* WT when the bacteria were challenged with an over-expression plasmid containing *cas1*, *cas2*, *csn2* and *cas9* as shown in our study (**Figure 12A**) and a previous study (Wei et al., 2015a). On the other hand, lytic phage challenge might confer higher selective pressure to the bacteria for triggering spacer acquisition. However, spacer acquisition with phage challenge was not examined in *S. pyogenes*, due to the unavailability of the lytic phage in our laboratory.

While spacer acquisition was not observed in the type II-A system of *S. pyogenes* via plasmid challenge approach, we detected active spacer acquisition in the endogenous type II-A system

of *S. thermophilus* LMD-9 via phage challenge and over-expression of Cas proteins. When *S. thermophilus* LMD-9 was challenged with lytic phage DT1, we observed an higher uptake of new spacers in CRISPR3 locus compared to CRISPR1 locus (**Figure 11; Table 1**), which could be explained by several possibilities, such as the PAM abundance in the phage genome. Based on phage challenge and *in silico* analyses, the PAM determined for CRISPR1 is 5′-NNAGAAW-3′ and for CRISPR3 is 5′-NGGNG-3′ (Horvath et al., 2008). The number of CRISPR3-associated PAM sequences in the genome of phage DT1 is twice as high as for CRISPR1-associated PAMs. Since PAMs are important for ensuring the acquisition of functional spacers, the higher amount of CRISPR3-associated PAMs in the genome of phage DT1 might favor spacer acquisition by CRISPR3 over CRISPR1. Furthermore, the higher spacer acquisition rate in the CRISPR3 locus may imply that the expression level the *cas* genes of the CRISPR3 locus of *S. thermophilus* LMD-9 is higher than those of the CRISPR1, as higher the expression levels of Cas proteins could enhance the acquisition activities (**Figure 12A**) (Wei et al., 2015a).

A very recent study revealed that AcrIIA6, an anti-type II-A CRISPR protein encoded by phage DT1, allows effective evasion of phage DT1 from CRISPR1 immunity but not from CRISPR3 immunity in a closely related strain, *S. thermophilus* SMQ-301 (Hynes et al., 2018). While the Cas9 inhibition mechanism of AcrIIA6 is still remained to be clarified (Hynes et al., 2018); another members of the type II-A anti-CRISPR family, AcrIIA2 and AcrIIA4, inhibit the *S. pyogenes* Cas9 activity by blocking the PAM-interacting site and the RuvC catalytic site (the blocking of RuvC was only shown in AcrIIA4) of Cas9, thereby precluding the target DNA recognition and endonuclease activity of Cas9 (Dong et al., 2017; Yang and Patel, 2017). Since the PAM-recognition and endonuclease activities of Cas9 are essential for interference, AcrIIA6 possibly inhibits Cas9 in a manner similar to AcrIIA2 and AcrIIA4. Assuming that this speculation is correct, the blocking of the PAM-interacting domain of Cas9 of the CRISPR1 locus by AcrIIA6 should not disturb spacer acquisition, however, it would lead to the acquisition of non-functional spacers (spacers without corresponding PAMs) in CRISPR1, as shown in a previous study (Heler et al., 2015). This would indicate that even though there is spacer acquisition in the CRISPR1 locus in the presence of AcrIIA6, the cells would still be killed by the lytic phage. With regard to this, the acquisition screening method to screen the survivors from the plaque assay could lead to bias results, which show low spacer acquisition rate in CRISPR1 locus. In this case, the acquisition-positive cells might have been killed by the lytic phages due to the acquisition of non-functional spacers and inhibition of the endonuclease

activity of Cas9 by AcrIIA6. Thereby, AcrIIA6 and survivor-based spacer acquisition screening method could be one of the reasons that we observed lower spacer acquisition in the CRISPR1 locus of *S. thermophilus* when it was challenged with lytic phage DT1 that encodes AcrIIA6.

Our spacer acquisition assay coupled with Cas proteins over-expression in *S. thermophilus* LMD-9 showed that Cas9 is essential for spacer acquisition, which is in line with the earlier studies (Heler et al., 2015; Wei et al., 2015a). When all type II-A Cas proteins from the CRISPR1 locus were over-expressed in the WT strain, spacer acquisition was detected in CRISPR1 locus (**Figure 12A**), but not CRISPR3 locus (**Figure 12B**). This suggests the absence evidence for the crosstalk between CRISPR1 and CRISPR3 for spacer acquisition. However, further experimental verification with the reciprocal over-expression of Cas proteins from the CRISPR3 locus and monitoring of the spacer acquisition in the CRISPR1 array is needed to confirm this. Additionally, it could be further verified by using a phage challenge approach to investigate spacer acquisition in the respective CRSIPR1 and CRISPR3 *cas* operon deletion mutants (the CRISPR arrays and leader sequences are not deleted). Since the leader-repeat boundary, especially LAS, is crucial for the recognition of Cas1-Cas2 for spacer integration (McGinn and Marraffini, 2016; Wei et al., 2015b; Wright and Doudna, 2016), crosstalk is unlikely due to the lack of sequence conservation on the leader-repeat boundaries and LAS between the CRISPR1 and the CRISPR3 as shown by our analyses (**Supplementary Figure S3**) and a literature (Van Orden et al., 2017). Furthermore, the lack of PPI between the Cas proteins from the CRISPR1 and the CRISPR3 in our pull-down assay also supports the unfeasibility of the crosstalk (**Figure 13B; Table 2**).

Altogether, we demonstrated that the two CRISPR loci of the type II-A systems of *S. thermophilus* LMD-9 are active in spacer acquisition and Cas9 is essential for acquisition. To gain more insights about whether non-CRISPR associated elements are involved in the spacer acquisition of the type II-A systems and to understand the interactions among the Cas proteins, we investigated the PPIs of Cas proteins.

4.2 Protein-protein interactions within and beyond the CRISPR-Cas systems

The association of Cas9, Cas1, Cas2 and Csn2 of *S. pyogenes* was demonstrated previously, however, the direct interactions between each of these components are still obscure (Heler et al., 2015). We addressed the direct interactions among the Cas proteins of the type II-A CRISPR-Cas systems via Y2H and pull-down assays in *S. pyogenes* and *S. thermophilus*, respectively. Pull-down assay revealed the direct interactions between Cas1 and Cas9 (prey is underlined) and Cas1-Cas1 dimer, whereas Y2H assay indicated the direct interactions between Cas1 and Csn2, Cas2 and Cas1, Csn2 and Cas9 and Cas1-Cas1 dimer. However, the identified interactions between two proteins are non-symmetric in Y2H assay, *i.e.* the interacting partner is not reversely detected when the fused domains of the bait and the prey are inverted, which could be explained by the fused domain-dependent steric hindrance – one of the limitations of Y2H assay (Brückner et al., 2009). The pull-down assay allows the detection of PPIs in their native conditions, which may be critical for some of the PPIs that might not be able to be detected in Y2H that lacks of the native cellular conditions (Brückner et al., 2009). This could explain the reason that some of the interacting partners detected in the pull-down assay were not observed in the Y2H assay.

The Cas1-Cas1 interaction showed by pull-down, Y2H (**Figure 13B**) and crosslinking assays (**Figure 18A**) confirmed that Cas1 forms a dimer, which is agreement with literature (Babu et al., 2011; Ka et al., 2016; Kim et al., 2013; Wiedenheft et al., 2009; Xiao et al., 2017). Moreover, we provided the evidence that the β-sheet-8 and the loop linking the β-sheet-8 and α-helix-2 are the possible Cas1 dimerization regions, via SPOT peptide assay and structural comparison between the Cas1 model of *S. thermophilus* with the Cas1 crystal structures of *S. pyogenes* and the Cas1-Cas2-prespacer complex of *E. faecalis* (**Figures 14B, 15A and 15C**). Some of the dimerization regions reported in the Cas1 of *S. pyogenes* (β-sheet-6, α-helix-1, loop linking the α-helix-6 and α-helix-7 and α-helix-5 to α-helix-8) (PDB: 4ZKJ) (Ka et al., 2016) and *E. faecalis* (β-sheet-6) (PDB: 5XVN) (Xiao et al., 2017), were not observed in our SPOT peptide assay, which indicates slight structural differences among the Cas1 orthologs. Although SPOT peptide assay could detect the sequence-dependent interaction, it has a limitation for the structure-dependent interaction, which could be one of the reasons that some

of the reported dimerization regions (Ka et al., 2016; Xiao et al., 2017) were not identified in our assay.

Noteworthy, pull-down assay indicated the interaction between Cas1 of the type II-A system and AddAB (**Figure 13B; Table 2**), which is consistent with the Cas1-RecBC interaction reported earlier in the type I-E system of *E. coli* (Babu et al., 2011). The homologous recombination complexes, AddAB and RecBCD, were shown to promote spacer acquisition in the type II-A and type I-E systems, respectively (Levy et al., 2015; Modell et al., 2017). This led to a proposal that Cas1-Cas2 complex captures the DNA degradation fragments generated by AddAB/RecBCD for spacer integration (Levy et al., 2015; Modell et al., 2017). Nonetheless, this model was challenged by a very recent study, which demonstrated that the helicase activity of RecBCD is required to promote spacer acquisition, instead of the nuclease activity (Radovčić et al., 2018). The authors further proposed that the helicase activity of the RecBCD facilitates spacer acquisition by removing the nucleoprotein complexes from the DNA damage sites, thereby allowing Cas1-Cas2 to access the DNA substrates (Radovčić et al., 2018). Based on our findings and literature, it is tempting to speculate that AddAB/RecBCD complex possibly recruits Cas1 to the DNA damage site, where Cas1 subsequently extracts the prespacer substrates via its nuclease activity. This hypothesis is supported by an *in vivo* study in the type I-E system of *E. coli* that showed the recruitment of Cas1 to the MMC-induced DNA DSBs, and an *in vitro* study that demonstrated that Cas1 cleaves the DNA intermediates of DNA repair and recombination pathways, such as replication fork, 5´-flap, 3´-flap and Holiday junction (Babu et al., 2011). To investigate whether AddAB/RecBCD recruits Cas1 to the DSB sites, *in vivo* study such as FRET, BRET and BiFC, or *in vitro* study such as an electrophoretic mobility shift assay (EMSA), could be applied.

The nuclease activity of Cas1 alone is not enough to ensure the acquisition of functional spacers that could provide immunity, and Cas9 is required for the PAM selection during the prespacer selection process to ensure the acquisition of functional spacers (Heler et al., 2015). In this work, our pull-down assay provided the evidence for a direct physical interaction between Cas1 and Cas9 for the first time (**Figure 13B; Table 2**). Our SPOT peptide assay (**Figure 14C**) and structural comparison (**Figures 15B-C**) further indicated that the β-sheet-10 and α-helix-10 on the outer surface of the C-terminus of Cas1 of *S. thermophilus* are possibly the interacting sites of Cas1 with Cas9, as these regions could be accessible by Cas9. We propose that Cas1 interacts with Cas9 during the prespacer selection process, and upon PAM recognition by Cas9, Cas1

extracts the corresponding prespacer with its nuclease activity. Cas1 could still extract prespacers from the DNA substrates when the PAM recognition function of Cas9 is deactivated, however, the presence of Cas9 is essential. This model is supported by an experiment showing that a double point mutation on the PAM-binding motif of Cas9 did not abolish spacer acquisition, nevertheless, it led to acquisition of non-functional spacer, *i.e.* spacers match to the protospacers without PAMs (Heler et al., 2015). This model could be investigated by examining the spacer acquisition *in vivo* in the mutants with the Cas9-Cas1 interacting regions and/or PAM-binding motif of Cas9 disrupted.

tracrRNA is essential for spacer acquisition, and the tracrRNA-bound Cas9 was suggested for exhibiting the proper conformation needed for spacer acquisition (Heler et al., 2015). It is worth mentioning that tracrRNA was present in our pull-down assay, therefore our pull-down approach provides the native conditions necessary for the direct interaction between Cas1 and tracrRNA-bound Cas9. The absence of tracrRNA in the Y2H assay could be a reason that Cas9-Cas1 interaction was not detected in the Y2H assay. While an earlier study showed the co-purification of all the four Cas proteins of the type II-A system of *S. pyogenes* by SEC (Heler et al., 2015), a more recent study could not show any evidence for the stable complex formation by the four Cas proteins of the type II-A system of *S. pyogenes*, neither the interactions between Cas9 and Cas1, Cas9 and His_6-MBP(maltose binding protein)-Cas2 or Cas9 and Cas1-His_6-MBP-Cas2 (Ka et al., 2018). Unlike our pull-down assay, tracrRNA was absent in the latter study (Ka et al., 2018) and our SEC studies (**Figure 17; Supplementary Figure S5**), which could be a possible reason that the Cas9-Cas1 interaction was not detected. Furthermore, the His_6-MBP tag that was used to solubilize Cas2 might hinder the interaction of Cas9 and Cas1 when the combination of Cas9 and Cas1-His_6-MBP-Cas2 was investigated via SEC (Ka et al., 2018).

Although tracrRNA was absent in the Y2H assay, the interaction between Csn2 and Cas9 protein was still detected (**Figure 13B; Supplementary Table S6**). This is consistent with a recent study demonstrating that both apo-Cas9 and sgRNA-bound Cas9 could interact with Csn2, however, the interaction that involves sgRNA-bound Cas9 is much stronger than apo-Cas9 (Ka et al., 2018). In addition to the Csn2-Cas9 interaction, the direct interaction between Cas1 and Csn2 was indicated in the Y2H assay (**Figure 13B; Supplementary Table S4**), which is in agreement with the previous studies (Ka et al., 2016, 2018). A model structure of Cas1-Cas2-Csn2 complex of the type II-A system of *S. pyogenes* generated by

in silico methods, indicates that a Csn2 tetramer and a Cas2 dimer, respectively, bind to of one of the N-terminal domains of a Cas1 dimer without overlapping (Ka et al., 2016). In line with the former studies (Ka et al., 2016, 2018), our findings showed the Cas1-Csn2 and Cas1-Cas2 interactions (**Figure 13B; Supplementary Tables S4 and S5**), but not the Cas2-Csn2 interaction. Although Cas2 does not interact directly with Csn2, the presence of Cas2 increases the binding affinity of Csn2 to Cas1 by approximately 10-fold (Ka et al., 2018). This indicates that Cas1-Cas2 complex possesses the appropriate conformation that facilitates the interaction between Csn2 and Cas1.

Both Csn2 and Cas9 are not required for spacer integration *in vitro*, therefore they were proposed for their roles in other step of spacer acquisition, for instance the generation of prespacers (Wright and Doudna, 2016). Based on the interactions of Csn2 with both Cas9 and Cas1 (**Figure 13B; Supplementary Tables S4 and S6**) (Ka et al., 2016, 2018), respectively, and the DNA end binding activity of Csn2 (Arslan et al., 2013), we suggest that Cas9 and/or Cas1 might facilitate the recruitment of Csn2 to the free DNA ends of the prespacer that was loaded on Cas1-Cas2 complex. This recruitment possibly enables Csn2 to protect the free DNA ends of the prespacer from the degradation by the cellular nucleases, which was suggested earlier in a study (Arslan et al., 2013). The physical presence of Cas9 is critical for spacer acquisition, as spacer acquisition was still detected when the catalytic domain and PAM-binding motif were mutated, respectively, although it led to acquisition of non-functional spacers in the latter case (Heler et al., 2015). This implies that in addition to the PAM recognition role, the physical presence of Cas9 prior to spacer integration has another role, such as to facilitate the recruitment of Csn2 to the DNA ends of the prespacer either alone or together with Cas1. Previously, the binding activity of Csn2 on the DNA ends of the radiolabeled DNA fragment was demonstrated by EMSA (Arslan et al., 2013). However, this assay did not study whether Csn2 can bind to the ends of the DNA fragment (prespacer) that has been loaded on the Cas1-Cas2 complex. To investigate the binding activity of Csn2 on the DNA ends of the prespacer that has been loaded on the Cas1-Cas2 complex, similar EMSA (Arslan et al., 2013) could be performed by using a DNA fragment that was pre-bound to the Cas1-Cas2 complex with and without the presence of Cas9.

In addition to the interactions among the Cas proteins, our work provides the indications of the interactions between the Cas proteins of the type II-A systems and DNA repair proteins (**Figure 13B; Table 2; Supplementary Tables S5-S8**), which included the interactions that

are supported by a previous study in *E. coli*, such as the interactions of Cas1 and AddAB (RecBC), and Cas1 and MutS (Babu et al., 2011). For the first time, we provide the evidence of the interactions between Cas proteins and the proteins from the NER, MMR and BER pathways, such as the interactions between Cas1 and UvrB, Cas2 and UvrA, Cas9 and UvrA, Cas2 and XseA, Cas9 and PcrA, and Csn2 and LigA. Among the interacting partners from the DNA repair pathways, UvrA and UvrB from the NER pathway are the most prominent candidates, because these two proteins are involved in the interactions with Cas1, Cas2 and Cas9 (**Figure 13B; Table 2; Supplementary Tables S5, S7 and S8**). UvrA and UvrB form a UvrAB complex surveillance scanning for a DNA lesion at the beginning of NER (Truglio et al., 2006; Van Houten et al., 2005). Therefore, Cas1, Cas2 and Cas9 are presumably interacting with the UvrAB complex, instead of the individual subunits.

In addition to the physical and genetic interactions, several *in vivo* studies in the type I systems revealed the association of the CRISPR-Cas systems with DNA repair and chromosome segregation (Babu et al., 2011; Hare et al., 2014; Williams et al., 2007). DNA repair pathways are conserved in both eukaryotes and prokaryotes, whereas CRISPR-Cas systems are only found in the sequenced genome of 90% of archaea and 50% of bacteria (Grissa et al., 2007; Makarova et al., 2015), the potential roles of the CRISPR-Cas systems in DNA repair remains to be clarified. On the other hand, there are increasing amount of studies revealing the involvement of DNA repair proteins in the CRISPR-Cas systems, especially in the type I-E systems (Babu et al., 2011; Ivancic-Bace et al., 2015; Levy et al., 2015; Modell et al., 2017; Radovčić et al., 2018). Altogether, our study provides the initial clues about the potential involvement of DNA repair proteins in the CRISPR-Cas systems and/or *vice versa*, for future detailed investigations.

5 Conclusion

This thesis shows active spacer acquisition in the type II-A systems of *S. thermophilus* LMD-9 via phage challenge and over-expression of Cas proteins, respectively. While lytic phage challenge confers the selective pressure to the bacteria for triggering spacer acquisition, over-expression of Cas proteins increases the spacer acquisition activity in the plasmid-based acquisition, which is otherwise more difficult to investigate under native conditions. When all Cas proteins from the CRISPR1 locus were over-expressed in *S. thermophilus* WT, crosstalk in spacer acquisition between the CRISPR1 and the CRISPR3 loci was not observed. Our protein-protein interaction studies provide the evidence of the direct physical interactions among numerous Cas proteins of the type II-A systems, including Cas9-Cas1, Cas1-Csn2, Cas1-Cas2 and Csn2-Cas9. This work reveals that Cas1 is a central player in spacer acquisition, because it interacts with all the other three Cas proteins, *i.e.* Cas1, Csn2 and Cas9, presumably to mediate the spacer acquisition mechanism. These findings regarding the direct Cas protein interactions provide the basis for further investigation of the biological significance of these interactions in the spacer acquisition mechanism of the type II-A systems, which require all the four Cas proteins. Moreover, the protein-protein interaction studies indicate the interactions between the Cas proteins and the DNA repair proteins, such as Cas1-AddA, Cas1-AddB, Cas1-MutS, Cas1-UvrB, Cas2-UvrA, Cas9-UvrA, Cas2-XseA, Cas9-PcrA and Csn2-LigA, of which some have not been reported before. These findings show the connection between CRISPR-Cas systems and DNA repair pathways, which is less known for the type II-A systems. Based on the findings in this thesis, future work shall focus on addressing the biological significances of the interactions among the Cas proteins of the type II-A systems. Moreover, the biological significance of the involvement of the DNA repair proteins in the mechanism of CRISPR-Cas systems and *vice versa*, will be novel fields for further exploration.

6 Materials and Methods

6.1 Bacterial strains and culture conditions

Bacterial strains used in this study are listed in the **Supplementary Table S10**. *S. pyogenes* was either grown in THY broth (Todd Hewitt Broth; Becton Dickinson) containing 0.2% yeast extract (Oxoid)) or on TSA plate (Trypticase^{TM} Soy agar; Becton Dickinson) supplemented with 3% sheep blood, at 37°C supplemented with 5% carbon dioxide (CO_2) without agitation. *S. thermophilus* was either grown in M17 broth or on M17 plate (Difco^{TM} M17 Broth) supplemented with 0.5% lactose (LM17), at 42°C supplemented with 5% CO_2 without agitation. For transformation purpose, *S. thermophilus* was either grown in chemically defined medium (CDM) with 1% lactose or on LM17 plate with 1% lactose. *E. coli* was grown in LB broth (with agitation) or agar at 37°C. Whenever necessary, appropriate antibiotics were applied in the broth or agar, *i.e.* final concentration of 300 µg/ml of kanamycin for *S. pyogenes*; 50 µg/ml of kanamycin, 5 µg/ml erythromycin, 2 µg/ml of chloramphenicol in LM17 broth (or 5 µg/ml chloramphenicol on LM17 plate) for *S. thermophilus;* and 25 µg/ml of kanamycin, 50 µg/ml of streptomycin, 5 µg/ml of tetracycline,100 µg/ml of ampicillin or 10 µg/ml of chloramphenicol for *E. coli*. The concentrations of the antibiotics differ from the ones mentioned here, will be indicated in the specific assays.

6.2 Bacterial transformation

Standard protocol was used for heat shock plasmid transformation into *E. coli* (Sambrook et al., 1989). Transformation of *S. pyogenes* was performed according to (Caparon and Scott, 1991). Natural transformation of *S. thermophilus* was conducted according to previous publication (Gardan et al., 2009) with some modifications. Briefly, overnight culture of *S. thermophilus* in CDM was diluted 1:100 in fresh CDM and grown until an OD_{600} (optical density at a wavelength of 600 nm) of 0.2 was reached. Five hundred µl of the culture was transferred to an Eppendorf tube that contained 200 ng plasmid DNA or PCR fragment and grown for 2 hours. Finally, the culture was plated on LM17 plate with appropriate antibiotics and grown overnight.

6.3 DNA manipulations

Standard protocols (Sambrook et al., 1989) were used for DNA manipulations, *i.e.* DNA preparation, PCR amplification, DNA digestion, ligation, purification, agarose gel electrophoresis. Genomic DNA preparation, plasmid DNA preparation and DNA purification were performed with kits (NucleoSpin Tissue Kit, Macherey-Nagel kit; Qiagen). Primers were provided by Sigma-Aldrich, and the primers were detailed in the **Supplementary Table S12**. PCR fragments and plasmid DNA were sequenced at either LGC Genomics or Seqlab GmbH to check for the insert sequences (for cloning) and acquired spacers (for spacer acquisition study).

6.4 Plasmid constructions for the heterologous type II-A system of *S. pyogenes*

Individual *S. pyogenes cas* genes or their combinations were PCR amplified and cloned into pCDF-DUET vector using BamHI and EcoRI restriction sites (**Supplementary Table S11-12**). TracrRNA-CRISPR and *cas9* of *S. pyogenes* were cloned into pEC85 vector using BamHI and EcoRI, and SmaI and SalI restriction sites, respectively. Protospacer-mutated-PAM (spy0700-TG(PAM)) was cloned into pUC19 vector using ZraI restriction sites.

6.5 RNA extraction

Bacterial culture of *E. coli* BL21-AI harboring pCDF-DUETΩ*cas* variants, was either not induced or induced with 0.5 mM IPTG and 0.2% arabinose for 2 hours, and then mixed with equal volume of ice-cold ethanol-acetone (1:1) solution. Total RNA was isolated using TRIzol (Invitrogen) reagent and chloroform, followed by isopropanol precipitation, and eventually purified with TurboDNase (Ambion), according to the manufacturer's instructions.

6.6 Semi-quantitative reverse transcription PCR (RT-PCR)

The DNase-free total RNAs were used as templates for RT-PCR, which was performed using Qiagen® OneStep RT-PCR Kit (Qiagen) according to the manufacturer's protocol and the primers are indicated in **Supplementary Table S12**. The absence of DNA contamination was verified with PCR using HotStarTaq Polymerase (Qiagen) following manufacturer's protocol.

6.7 Plasmid-based spacer acquisition study in the heterologous type II-A system of *S. pyogenes*

For two-plasmid system, the plasmids pCDF-DUETΩ*cas1cas2csn2* (pEC651) (streptomycin resistant) and pEC85ΩtracrRNA-Leader-CRISPR (pEC645) (kanamycin resistant) were stepwise transformed into the *E. coli* BL21-AI. A single colony harboring two plasmids was inoculated in 50 ml LB medium (streptomycin; kanamycin; 0.2% glucose) and grown over-day at 37˚C, which was subsequently diluted (1:250 dilution) in fresh LB medium (streptomycin; kanamycin; 0.2% glucose) and grown overnight at 37˚C. The next day, the overnight culture was diluted (1:250 dilution) and grown over-day again at 37˚C. The over-day culture was subsequently diluted (1:250 dilution) in two different sets of LB medium, *i.e.* (1) without antibiotics (0.2% arabinose and 0.1 mM IPTG) and another with antibiotics (streptomycin; kanamycin; 0.2% arabinose; 0.5 mM IPTG). At this point, 0.2% arabinose and 0.1 mM IPTG was added to both LB media to induce the expression of *cas* genes that were controlled under the T7-*lac* promoter. These media were incubated at 37˚C overnight. Afterwards, the 2 sets of culture were diluted with fresh LB medium with and without antibiotics, respectively, twice a day (over-day and overnight) and grown at 37˚C for a long period of time (~1 to 2 weeks). At the same time, an aliquot of the same overnight culture was also diluted in 50 ml LB medium (1:250 dilution) with 25 μg/ml kanamycin and 0.2% glucose to suppress the *cas* expression for sampling purpose. The bacterial culture was collected for every cycle of the dilution, followed by plasmid DNA purification, genomic DNA purification and PCR analysis. Each inoculation and sub-growing was named as one cycle. For spacer acquisition screening, the CRISPR arrays of *S. pyogenes* and *E. coli* BL21-AI were monitored by PCR by using primers listed in **Supplementary Table S12**.

For three-plasmid system, the plasmids pCDF-DUETΩ*cas1cas2csn2* (pEC651), pEC85ΩtracrRNA-Leader-CRISPR-*cas9* (pEC663) and pUC19Ωspy0700-TG(PAM) (pEC687) (ampicillin resistant) were stepwise transformed into the *E. coli* BL21-AI. A single colony was inoculated in 10 ml LB medium (streptomycin; kanamycin and ampicillin) and incubated at 37˚C overnight. The overnight culture was diluted (1:50 dilution) in 25 ml LB medium (streptomycin; kanamycin and ampicillin) and incubated at 37˚C until an OD_{600} of approximately 0.5 was achieved. At this point, 1 ml of the culture was centrifuged and the pellet was resuspended with 1 ml of fresh LB medium, wherein 100 µl of the resuspension was inoculated in 25 ml of LB medium (1:250 dilution) (streptomycin; kanamycin and ampicillin), and the *cas* expression was induced with 0.2% arabinose and 0.5 mM IPTG. The culture was incubated at 37˚C overnight, and the overnight culture was diluted (1:250 dilution) in fresh LB medium (streptomycin; kanamycin; 0.2% arabinose; 0.5 mM IPTG) and grown at 37˚C over-day. These over-day and overnight dilution and growing steps were repeated for a long period of time (~1 to 2 weeks). The culture was plated on LB agar plate (streptomycin; kanamycin; 0.2% arabinose and 0.5 mM IPTG) and incubated at 37˚C over-day or overnight. Later, 100 colonies from this plate were replica plated on two sets of plates (streptomycin; kanamycin; 0.2% arabinose and 0.5 mM IPTG), *i.e.* one plate without ampicillin and another plate with ampicillin. The ampicillin sensitive colonies were analyzed by colony PCR (**Supplementary Table S12**) to check for the expansion of CRISPR array (*i.e.* spacer acquisition)

6.8 Spacer acquisition study in the heterologous type II-A system of *S. pyogenes* via phage challenge assay

pCDF-DUETΩ*cas1cas2csn2* (pEC651) (streptomycin resistant) and pEC85ΩtracrRNA-Leader-CRISPR_*Cas9* (pEC663) (kanamycin resistant) were stepwise transformed into the *E. coli* BL21-AI. From the transformation, a single colony was inoculated into 5 ml of LB medium (10 mM $MgSO_4$; 0.2% maltose; 0.2% glucose; kanamycin; streptomycin) and incubated at 37°C overnight with agitation. The overnight pre-culture was diluted (1:50) in 10 ml LB medium (10 mM $MgSO_4$; 0.2% maltose; 0.2% glucose; kanamycin; streptomycin), and then grown until an OD_{600} of 0.6-0.8 was achieved. The culture was

centrifuged at 4,000 rpm for 10 minutes at 4°C, and the pellet was re-suspended with equal amount of fresh LB medium (10 mM $MgSO_4$; 0.2% maltose; kanamycin; streptomycin). The expression of *cas* genes were induced by adding 0.5 mM IPTG and 0.2% arabinose, and the culture was grown for 2 hours at 37°C with agitation. Afterwards, the culture was diluted 10 times and grown overnight. 100 µl of overnight culture was mixed with 100 µl of phage λ_{vir} (approximate 10^6 to 10^{10} PFU/ml of phage λ_{vir} in 100 µl of SM medium (100 mM NaCl; 8 mM $MgSO_4 \bullet 7H_2O$; 50 mM Tris-HCl, pH7.5) depending on the MOI), and the mixture was incubated at 37°C for 60 minutes without agitation, followed by an incubation at 37°C for 45 minutes with agitation. 4 ml LB medium (10 mM $MgSO_4$; 0.2% maltose; 0.5 mM IPTG; 0.2% arabinose) was added to the infected culture and grown overnight. The overnight culture was plated on LB agar plate (0.5 mM IPTG; 0.2% arabinose; kanamycin) and incubated at 37°C for 3 days. The colonies were screened for spacer acquisition using PCR.

6.9 Plasmid-based spacer acquisition study in the endogenous type II-A system of *S. pyogenes*

For acquisition study in *S. pyogenes* WT strain (EC904), pEC85 vector (kanamycin resistant) was transformed into WT via electroporation and plated on TSA plate supplemented with 300 µg/ml kanamycin. Resulting colonies were inoculated in 25 ml of THY medium (kanamycin) and grown over-day. The over-day pre-culture was diluted 1:250 in fresh THY medium with (kanamycin) and grown overnight. For the initial two steps, the kanamycin was present in the medium to ensure a stable environment for the maintenance of the plasmid. Thereafter, the culture was diluted (1:250 dilution) twice a day in THY medium without antibiotics for a period of at least two weeks. If spacer acquisition occurs, it will lead to interference, *i.e.* plasmid loss. Here, to assure the survival of the cells that have lost the antibiotic resistance due to plasmid loss, THY medium without antibiotics was used. The first batch of bacteria that grew in the medium without antibiotics is known as cycle 1, the culture of subsequent dilution as cycle 2 and so on. An aliquot of the culture was sampled for every cycle and spread on the TSA plates without antibiotics, and then replica plated on plates with and without kanamycin on the next day. Genomic DNA was extracted from the colonies that only grew on plates without antibiotics, and then it was used as a DNA template for PCR screening for spacer acquisition.

For the acquisition study in *S. pyogenes* Δ*cas9* strain (EC1788), pEC85, pEC659 and pEC660 were transformed into the Δ*cas9* strain EC1788 via electroporation and plated on the TSA plate supplemented with kanamycin. Resulting colonies were inoculated in 25 ml of THY medium supplemented with kanamycin and grown overnight. The overnight culture was diluted 1:250 in fresh THY medium with kanamycin on the next morning, and it was subsequently grown overnight. The overnight culture was repeatedly diluted and grown on a daily basis for a duration of at least two weeks. An aliquot of the culture from every cycle was sampled for genomic DNA extraction and then used as a DNA template for PCR screening for spacer acquisition. PCR products were analyzed on 1.5% agarose gel (1x TBE) and visualized by staining with ethidium bromide.

6.10 PCR analysis for spacer acquisition

The template DNA for PCR was either 10 ng plasmid DNA, 50 ng genomic DNA, bacterial colony or culture. For colony PCR, a single colony was transferred to 15 µl sterile water and boiled at 95°C for 10 minutes. For bacterial culture-based PCR, 1 ml of the bacterial culture was centrifuged, the pellet was washed in 1 ml of sterile water, re-suspended with 500 µl of sterile water, and boiled at 95°C for 10 minutes. For both colony and bacterial culture-based PCR, 5 µl bacterial suspension was used as a template for the PCR reaction. The PCR reaction was prepared with a final concentration of 0.10 U/µl of Taq DNA polymerase (Thermo Scientific), 1x Taq buffer with $(NH_4)_2SO_4$, 2 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 µM forward and reversed primers. The PCR cycling condition was: initial denaturation at 95˚C for 5 minutes, denaturation at 95˚C for 30 seconds, annealing at 54˚C for 30 seconds, extension at 72˚C for 40 seconds (35 cycles for denaturation to extension) and a final extension at 72˚C for 7 minutes. The primers used are indicated in the **Supplementary Table S12**. The PCR products were analyzed by agarose gel electrophoresis.

6.11 Plasmid constructions for the heterologous type II-A system of *S. thermophilus*

Individual *S. thermophilus cas* genes were PCR amplified and cloned into vector pCDF-DUET using BamHI and NcoI restriction sites (**Supplementary Table S11-12**). The CRISPR array was cloned into vector pUC85 using BamHI and EcoRI restriction sites.

6.12 Plasmid-based spacer acquisition study in the heterologous type II-A system of *S. thermophilus*

The *S. thermophilus* heterologous plasmids pCDF-DUETΩ*cas1-cas2-csn*2 (pEC1151) (streptomycin resistant) and pEC85ΩLeader-CRISPR1 (pEC1230) (kanamycin resistant) were stepwise transformed into the *E. coli* BL21-AI host. A single colony was inoculated in 20 ml LB medium (50 µg/ml streptomycin; 25 µg/ml kanamycin; 0.2% glucose) in the morning and grown at 37°C (with agitation) over-day. One ml of the over-day grown culture was sampled for PCR screening for spacer acquisition, after washing in 1x PBS buffer twice. The same culture was also diluted (1:500) in 20 ml fresh LB medium (25 µg/ml kanamycin; 0.1 mM IPTG; 0.2% glucose) and grown overnight at 37°C (with agitation). Starting from this step, the bacterial culture was diluted (1:250) twice a day until cycle 24, and the cultures of each cycle were sampled for PCR screening. On cycle 16, an aliquot of the culture was diluted and spread on a LB agar plate supplemented with kanamycin (25 µg/ml). On the next day, the colonies were replica plated on LB agar plate with (1) streptomycin and kanamycin, and with (2) kanamycin only. A number of the colonies that grew only on the kanamycin plates (without streptomycin) were checked for spacer acquisition by colony PCR.

6.13 Spacer acquisition study in the heterologous type II-A system of *S. thermophilus* via phage challenge assay

E. coli BL21-AI harboring plasmids pCDF-DUETΩ*cas1-cas2-csn*2 (pEC1151) (streptomycin resistant) and pEC85ΩLeader-CRISPR1 (pEC1230) (kanamycin resistant) was studied here. A single colony was inoculated in 50 ml LB medium (50 µg/ml streptomycin and 25 µg/ml kanamycin) and grown at 37°C (with agitation) overnight. The pre-culture was diluted (1:100) and induced in 200 ml of fresh LB medium (50 µg/ml streptomycin; 25 µg/ml kanamycin; 0.1 mM IPTG; 0.2% arabinose; 10 mM $MgCl_2$; 5 mM $CaCl_2$; 0.2% maltose) and grown until an OD_{600} of 0.4-0.5 was reached. Equal amounts of the bacterial culture and phage λ_{vir} were mixed, wherein the concentration of the phage was adjusted according to the MOIs used, namely MOIs of 0.1, 1 and 10. The infected cultures with different MOIs were incubated at

37°C without agitation, for 5, 10, 15 and 20 minutes, respectively, followed by 15 minutes of incubation at 37°C with agitation. The bacteria-phage mixture was subsequently mixed with 0.7% soft agarose and plated on a LB agar plate (25 μg/ml kanamycin; 0.1 mM IPTG; 0.2% arabinose; 10 mM $MgCl_2$; 5 mM $CaCl_2$; 0.2% maltose). The plate was incubated at 37°C for 1-3 days, and the colonies were screened for spacer acquisition.

6.14 Spacer acquisition study of the endogenous type II-A systems of *S. thermophilus* via phage challenge assay

An overnight culture of *S. thermophilus* was diluted in 10 ml of LM17 medium and grown until an OD_{600} of approximately 0.5 was reached. $CaCl_2$ (10 mM) was added to the culture. Five hundred μl of the culture was mixed with an appropriate amount of phage DT1 to achieve the desired MOI. The bacteria-phage mixture was then quickly mixed with 0.7% soft agarose supplemented with 0.5% lactose and 10 mM $CaCl_2$, and subsequently transferred to a LM17 plate, which was incubated at 42°C for 1-3 days. The colonies were screened for spacer acquisition.

6.15 Spacer acquisition study of the endogenous type II-A systems of *S. thermophilus* with Cas proteins over-expression

Plasmid pCas1-Cas2-Csn2-Cas9 or pCas1-Cas2-Csn2 (gifts from Michael Terns) was transformed into *S. thermophilus* and plated on LM17 plate (5 μg/ml chloramphenicol). A single colony was inoculated in 5 ml of LM17 medium (2 μg/ml chloramphenicol) and grown overnight. One ml of the overnight culture was centrifuged, and the resulting pellet was washed with sterile water and centrifuged again. Afterward, the pellet was re-suspended with sterile water and boiled at 95°C for 10 minutes. Five μl of the bacterial suspension was used as DNA template for PCR screening for spacer acquisition. The PCR product was analyzed with 2% TAE-agarose gel stained with ethidium bromide.

6.16*In vitro* pull-down assay

For cell lysate preparation, an *S. thermophilus* overnight culture was diluted in LM17 medium and grown until an OD_{600} of approximately 0.70 was reached. The culture was harvested by centrifugation at 6,000 rpm for 20 minutes at 4°C, and the pellet was subsequently washed with 1 ml of 1x PBS buffer twice. The washed pellet was dissolved in thw lysis buffer (50 mM Tris-HCl, pH 7.5; 150 mM NaCl; 0.1% Triton X-100) and lysed with FastPrep (MP Biomedicals). Next, the cell lysate was centrifuged at 13,000 rpm for 10 minutes at 4°C. The concentration of the total protein (in the supernatant) of the cell lysate was determined by standard Bradford assay.

For the *in vitro* pull-down, 150 µg of the purified CPD-His_{12}-tagged Cas1 (the bait) was pre-incubated with 300 µg of total cell lysate (containing the prey) at room temperature for 40 minutes with 100 rpm rotation. Afterward, the bait-prey sample was incubated with TALON Co^{2+} resin (GE Healthcare) at room temperature for 20 minutes with 150 rpm rotation. The sample was centrifuged briefly, and the collected solution was the flowthrough sample. The mini column was washed the wash buffers (10 mM Tris-HCl, pH8; 300 mM NaCl; 5 mM β-Mercaptoetanol; 10% Glycerol) containing 20 mM imidazole for wash 1-3 and 50 mM imidazole for wash 4-5. Next, 50 µM phytic acid buffer was incubated with the samples at room temperature for 1 hour with 150 rpm rotation. Phytic acid buffer was used to detach the CPD tag from Cas1 protein, which was eluted from the column. The elution was analyzed on Coomassie stained-gradient SDS-PAGE gel. The control (cell lysate only) was prepared accordingly without the addition of the bait, and same for the preparation of bait only control, which was not incubated with the cell lysate. Eventually, all the bait-prey samples, cell lysate only control and bait only control were sent for mass-spectrometry analysis to identify the interacting partners of Cas1.

6.17SPOT peptide assay

The Cas1 SPOT peptide membrane was blocked with blocking buffer (5% skim milk powder (Sigma-Aldrich) in 1x TBS-T buffer) at 4°C overnight. Next day, the membrane was washed 3 times with 1x TBS-T at 4°C for 5 minutes, followed by the washing with protein storage buffer (buffer for Cas1: 50 mM Tris-HCl; 500 mM NaCl; 20% glycerol; 5 mM β-mercaptoethanol; buffer for Cas9: 20 mM HEPES, pH7.5; 150 mM KCl; 20% glycerol) at 4°C for 10 minutes.

Afterwards, the membrane was incubated with 50 nM His_6-tagged Cas1 or His_6-tagged Cas9 pre-incubated with tracrRNA, in fresh protein storage buffer at 4°C for 1.5 hours. His_6-tagged Cas9 was pre-incubated with two molar excess of tracrRNA at 37°C for 10 minutes. Subsequently, the membrane was washed with protein storage buffer at 4°C for 5 minutes, followed by 1x TBS-T buffer at 4°C for 5 minutes for 3 times. Next, the membrane was incubated with Penta-His antibody (mouse monoclonal IgG1) (Qiagen) in 1x TBS-T buffer (1:2,000 dilution) at room temperature for 1.5 hours, followed by multiple washing in 1x TBS-T buffer. The membrane was subsequently incubated with Mouse IgG HRP (horseradish peroxidase-conjugated) Linked Whole Ab (GE Healthcare) (1:10,000 dilution) in blocking buffer for 1.5 hour. After the incubation with the secondary antibody, the membrane was washed with 1x TBS-T buffer for 5 mins at room temperature for several times. Finally, the SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo Scientific) was added to the SPOT peptide membrane and analyzed by ChemiDoc™ Gel Imaging System (Bio-Rad).

6.18 Protein purification

The expression strain *E. coli* BL21 (DE3) was used for expressing CPD-His_{12}-tagged Cas1 (C-terminal), His_6-tagged Cas1 (C-terminal) and SUMO-His_6-tag Cas2 (N-terminal); whereas *E. coli* NiCo21(DE3) was used for expressing His_6-tagged Cas9 (C-terminal). The pre-culture of an expressing strain harboring a specific expression plasmid was diluted in LB medium (ampicillin) and grown until an OD_{600} of 0.6-0.8 was reached. The expression of Cas proteins was induced with 0.5 mM IPTG at either 18˚C (*cas1*) or 13°C (*cas9*) overnight, or 37°C (*cas2*) for 4 hours. The culture was harvested at 6,000 rpm, 4˚C for 15 minutes. The resulting pellet was re-suspended and lysed in lysis buffer, and harvested again at 18,000 rpm, 4˚C for 45 minutes. The clarified lysate was bound to TALON Co^{2+} resin (GE Healthcare) and was washed with wash buffer. After extensive washing, the protein was eluted with elution buffer and analyzed on Coomassie stained-SDS-PAGE gel.

After IMAC purification with TALON Co^{2+} resin, His_6-tagged Cas1 or CPD-His_{12}-tagged Cas1 was further purified with SEC. SUMO protease was incubated with SUMO-His_6-tag Cas2 at 4°C for 1 hour to remove the SUMO-tag before SEC. The TALON-purified His_6-tagged Cas9 was further purified with chitin resin (NEB) before SEC. Here, His_6-tagged Cas9 was incubated with chitin resin at 4°C for 1 hour, and subsequently eluted with buffer A. The SEC column

used for protein purification was HiLoad® 16/600 Superdex® 200 pg (GE Healthcare). The eluted protein was analyzed again on Coomassie stained-SDS-PAGE gel.

For *in vitro* pull-down assay (CPD-His_{12}-tagged Cas1 was used) and SPOT peptide assay (His_6-tagged Cas1 was used), the recombinant Cas1 were purified with Tris-HCl-based lysis/wash buffers, whereas His_6-tagged Cas9 was purified in HEPES-KOH-based buffers (**Supplementary Table S9**). For SEC and crosslinking assays, HEPES-HCl-based buffers with lower pH were used for the purification of His_6-tagged Cas1, SUMO-His_6-tag Cas2 and His_6-tagged Cas9, to improve the protein stability

6.19 Protein-protein interaction study via size-exclusion chromatography

With regard to the study of Cas9 and Cas1 complex formation, the purified Cas9 and Cas1 (see the protocols for protein purification) were pooled together and incubated at 4°C for 45 minutes, and their complex formation was subsequently purified with an analytical column – Superdex® 200 Increase 10/300 GL by SEC using the buffer described in **Supplementary Table S9**. The elutions were analyzed on Coomassie stained-SDS-PAGE gel. To compare elution profile of Cas9-Cas1 with Cas1, Cas1 was also individually purified on Superdex® 200 Increase 10/300 GL column, and the elutions were analyzed on Coomassie stained-SDS-PAGE gel.

As for the study of Cas9-Cas1-Cas2 complex formation, each recombinant protein was expressed individually, and then their cell lysates were combined and purified via IMAC and SEC as described in the protocol for protein purification. SUMO protease was added to the purified proteins mixture to cleave off the SUMO-His_6-tag of Cas2 before the subsequent SEC purification with Superdex® 200 Increase 10/300 GL. The eluates from the SEC were analyzed on Coomassie stained-SDS-PAGE gel, to verify their complex formation.

6.20 Crosslinking

Different combinations of Cas proteins were pre-incubated together at 4°C for 1 hour in protein buffer (20 mM HEPES-HCl, pH 7.0; 250 mM NaCl; 10% glycerol; 0.05% Tween-20).

His_6-tagged Cas9 was pre-incubated with 2 times molar excess of tracrRNA at 37°C for 10 minutes, before the incubation with other Cas protein(s). To crosslink the interacting proteins, BS^3 crosslinker was added to the mixture and incubated at 24°C for 30 minutes. The crosslinking reaction was quenched by the addition of Tris-HCl (50 mM), followed by 15 minutes incubation at 24°C. Then, the reaction was boiled at 95°C for 5 minutes and analyzed on Coomassie-stained SDS-PAGE gel (10% or 12%), 4-20% Mini-PROTEAN®TGX TM Precast Gel (Biorad) or by Western Blot.

6.21 Western Blot

Crosslinking reactions were separated by 4-20% Mini-PROTEAN® TGX TM Precast Gel (Biorad). The proteins were transferred to a nitrocellulose membrane (GE Healthcare) via Mini Trans-Blot® Cell (Biorad). The membrane was subsequently blocked with blocking buffer [5% skim milk powder (Sigma-Aldrich) in 1x TBS-T buffer] at room temperature for 1 hour, followed by incubation with anti-Cas1, anti-Cas2 or anti-Cas9 antibody (1:1,000 dilution, BioGenes) in 1x TBS-T buffer with 2.5% skim milk powder, at 4°C for 2 hours. After extensive washing with 1x TBS-T buffer, the membrane was incubated with 1:10,000 dilution of anti-rabbit IgG HRP-linked secondary antibody in 1x TBS-T buffer, at room temperature for 1 hour. The membrane was washed extensively with 1x TBS-T buffer. Finally, SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo Scientific) was added to the membrane, and the signals were visualized with ChemiDoc™ Gel Imaging System (Bio-Rad).

7 Work Contributions

Unless otherwise mentioned, all works were performed by Shi Pey Wong.

Hagen Richter[2,3], Frank Hille [2,3,4,5] and Shi Pey Wong[1,2,3,4,5] conducted the endogenous *S. thermophilus* spacer acquisition study via phage challenge.

Shi Pey Wong[1,2,3,4,5], Katja Schmidt[2,3,4], Hagen Richter[2,3] and Ann Kathrin Ahrens[3] generated the *S. thermophilus cas* or CRISPR loci knock-out mutants.

Shi Pey Wong[1,2,3,4,5] and Katja Schmidt[2,3,4] purified the *S. thermophilus* Cas proteins.

Ines Fonfara[1,2,3] coordinated with the samples arrangement with Hybrigenics Services and generated the model structure of *S. thermophilus* LMD-9 Cas1.

Hybrigenics Services[6] provided the services of yeast two-hybrid screening.

The laboratory Henning Urlaub[7] provided mass-spectrometry services for checking the protein samples obtained from *S. thermophilus* Cas1 *in vitro* pull-down.

Peter Jungblut[8] provided mass-spectrometry services for checking the protein samples obtained from size-exclusion chromatography.

Affiliation

[1]The Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå Centre for Microbial Research (UCMR), Department of Molecular Biology, Umeå University, Umeå 90187, Sweden.

[2]Helmholtz Centre for Infection Research, Department of Regulation in Infection Biology, Braunschweig 38124, Germany.

[3]Max Planck Institute for Infection Biology, Department of Regulation in Infection Biology, Berlin 10117, Germany.

[4]Max Planck Unit for the Science of Pathogens, D-10117 Berlin, Germany.

[5]Institute for Biology, Humboldt University, 10115 Berlin, Germany

[6]Hybrigenics Services, 3 Impasse Reille, 75014 Paris, France

[7]Bioanalytical Mass Spectrometry Group, Max Planck Institute for Biophysical Chemistry, D-37077 Göttingen, Germany. Bioanalytics Group, Institute for Clinical Chemistry, University Medical Center Göttingen, D-37075 Göttingen, Germany.

[8]Core Facility Protein Analysis, Max-Planck-Institute for Infection Biology, Berlin 10117, Germany.

8 Supplementary Figures

A

```
SF370              1 ----tactc-ttaata-aatgca------------gtaatac-----ag-     26
                       |||||  ||||.|  ||||.|            ||.||||     ||
BL21-AI            1 taagtactctttaacataatggatgtgttgtttgtgtgatactataaagt     50

SF370             27 -gg--gcttttcaagactgaagtctagctgagacaaatagtgcgattacg     73
                      ||   |.||.|    |||||       |||.| |.||||     .|||
BL21-AI           51 tggtagattgt---gactg-------gctta-aaaaat-----catt---     81

SF370             74 aaatttttagacaaaaatagtctacg---ag-    102
                      ||||        ||.||||||..||.|   ||
BL21-AI           82 -aatt--------aataataggttatgtttaga    105
```

B

```
SF370              1 gttttagagctatgctgttttgaatggtccc--------------aaaac     36
                                           ||   |.||||              .||||
BL21-AI            1 ----------------------GA--GTTCCCCGCGCCAGCGGGGATAAAC     27

SF370             37 --     36

BL21-AI           28 CG     29
```

Supplementary Figure S1. Sequence similarity of the leader and repeat sequences between the type II-A system of *S. pyogenes* SF370 and the type I-E system of *E. coli* BL21-AI.

(A) The pairwise sequence alignments showed that the leader sequences between *S. pyogenes* SF370 and *E. coli* BL21-AI share only 46.6% of identity. The leader-anchoring sequence (LAS) of *S. pyogenes* is highlighted in grey. **(B)** The sequence identity of the repeat sequences between *S. pyogenes* SF370 and *E. coli* BL21-AI is only 21.2%. Pairwise sequence alignment (EMBOSS Needle) was used for analyzing the sequence similarity.

A

```
LMD-9_Cas9_Cr      1 MSDLVLGLDIGIGSVGVGILNKVTGEIIHKNSRIFPAAQAENNLVRRTNR     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1      1 MSDLVLGLDIGIGSVGVGILNKVTGEIIHKNSRIFPAAQAENNLVRRTNR     50

LMD-9_Cas9_Cr     51 QGRRLARRKKHRRVRLNRLFEESGLITDFTKISINLNPYQLRVKGLTDEL    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1     51 QGRRLTRRKKHRRVRLNRLFEESGLITDFTKISINLNPYQLRVKGLTDEL    100

LMD-9_Cas9_Cr    101 SNEELFIALKNMVKHRGISYLDDASDDGNSSVGDYAQIVKENSKQLETKT    150
                     ||||||||||||||||||||||||||||||||:|||||||||||||||||||
DGCC_Cas9_Cr1    101 SNEELFIALKNMVKHRGISYLDDASDDGNSSIGDYAQIVKENSKQLETKT    150

LMD-9_Cas9_Cr    151 PGQIQLERYQTYGQLRGDFTVEKDGKKHRLINVFPTSAYRSEALRILQTQ    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    151 PGQIQLERYQTYGQLRGDFTVEKDGKKHRLINVFPTSAYRSEALRILQTQ    200

LMD-9_Cas9_Cr    201 QEFNPQITDEFINRYLEILTGKRKYYHGPGNEKSRTDYGRYRTSGETLDN    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    201 QEFNPQITDEFINRYLEILTGKRKYYHGPGNEKSRTDYGRYRTSGETLDN    250

LMD-9_Cas9_Cr    251 IFGILIGKCTFYPDEFRAAKASYTAQEFNLLNDLNNLTVPTETKKLSKEQ    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    251 IFGILIGKCTFYPDEFRAAKASYTAQEFNLLNDLNNLTVPTETKKLSKEQ    300

LMD-9_Cas9_Cr    301 KNQIINYVKNEKAMGPAKLFKYIAKLLSCDVADIKGYRIDKSGKAEIHTF    350
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC Cas9 Cr1    301 KNQIINYVKNEKAMGPAKLFKYIAKLLSCDVADIKGYRIDKSGKAEIHTF    350

LMD-9_Cas9_Cr    351 EAYRKMKTLETLDIEQMDRETLDKLAYVLTLNTEREGIQEALEHEFADGS    400
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    351 EAYRKMKTLETLDIEQMDRETLDKLAYVLTLNTEREGIQEALEHEFADGS    400

LMD-9_Cas9_Cr    401 FSQKQVDELVQFRKANSSIFGKGWHNFSVKLMMELIPELYETSEEQMTIL    450
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    401 FSQKQVDELVQFRKANSSIFGKGWHNFSVKLMMELIPELYETSEEQMTIL    450

LMD-9_Cas9_Cr    451 TRLGKQKTTSSSNKTKYIDEKLLTEEIYNPVVAKSVRQAIKIVNAAIKEY    500
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    451 TRLGKQKTTSSSNKTKYIDEKLLTEEIYNPVVAKSVRQAIKIVNAAIKEY    500

LMD-9_Cas9_Cr    501 GDFDNIVIEMARETNEDDEKKAIQKIQKANKDEKDAAMLKAANQYNGKAE    550
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    501 GDFDNIVIEMARETNEDDEKKAIQKIQKANKDEKDAAMLKAANQYNGKAE    550

LMD-9_Cas9_Cr    551 LPHSVFHGHKQLATKIRLWHQQGERCLYTGKTISIHDLINNSNQFEVDHI    600
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    551 LPHSVFHGHKQLATKIRLWHQQGERCLYTGKTISIHDLINNSNQFEVDHI    600

LMD-9_Cas9_Cr    601 LPLSITFDDSLANKVLVYATANQEKGQRTPYQALDSMDDAWSFRELKAFV    650
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    601 LPLSITFDDSLANKVLVYATANQEKGQRTPYQALDSMDDAWSFRELKAFV    650

LMD-9_Cas9_Cr    651 RESKTLSNKKKEYLLTEEDISKFDVRKKFIERNLVDTRYASRVVLNALQE    700
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    651 RESKTLSNKKKEYLLTEEDISKFDVRKKFIERNLVDTRYASRVVLNALQE    700

LMD-9_Cas9_Cr    701 HFRAHKIDTKVSVVRGQFTSQLRRHWGIEKTRDTYHHHAVDALIIAASSQ    750
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    701 HFRAHKIDTKVSVVRGQFTSQLRRHWGIEKTRDTYHHHAVDALIIAASSQ    750

LMD-9_Cas9_Cr    751 LNLWKKQKNTLVSYSEDQLLDIETGELISDDEYKESVFKAPYQHFVDTLK    800
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    751 LNLWKKQKNTLVSYSEDQLLDIETGELISDDEYKESVFKAPYQHFVDTLK    800

LMD-9_Cas9_Cr    801 SKEFEDSILFSYQVDSKFNRKISDATIYATRQAKVGKDKADETYVLGKIK    850
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    801 SKEFEDSILFSYQVDSKFNRKISDATIYATRQAKVGKDKADETYVLGKIK    850

LMD-9_Cas9_Cr    851 DIYTQDGYDAFMKIYKKDKSKFLMYRHDPQTFEKVIEPILENYPNKQINE    900
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    851 DIYTQDGYDAFMKIYKKDKSKFLMYRHDPQTFEKVIEPILENYPNKQINE    900

LMD-9_Cas9_Cr    901 KGKEVPCNPFLKYKEEHGYIRKYSKKGNGPEIKSLKYYDSKLGNHIDITP    950
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    901 KGKEVPCNPFLKYKEEHGYIRKYSKKGNGPEIKSLKYYDSKLGNHIDITP    950

LMD-9_Cas9_Cr    951 KDSNNKVVLQSVSPWRADVYFNKTTGKYEILGLKYADLQFEKGTGTYKIS   1000
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1    951 KDSNNKVVLQSVSPWRADVYFNKTTGKYEILGLKYADLQFEKGTGTYKIS   1000

LMD-9_Cas9_Cr   1001 QEKYNDIKKKEGVDSDSEFKFTLYKNDLLLVKDTETKEQQLFRFLSRTMP   1050
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1   1001 QEKYNDIKKKEGVDSDSEFKFTLYKNDLLLVKDTETKEQQLFRFLSRTMP   1050

LMD-9_Cas9_Cr   1051 KQKHYVELKPYDKQKFEGGEALIKVLGNVANSGQCKKGLGKSNISIYKVR   1100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas9_Cr1   1051 KQKHYVELKPYDKQKFEGGEALIKVLGNVANSGQCKKGLGKSNISIYKVR   1100

LMD-9_Cas9_Cr   1101 TDVLGNQHIIKNEGDKPKLDF   1121
                     |||||||||||||||||||||
DGCC_Cas9_Cr1   1101 TDVLGNQHIIKNEGDKPKLDF   1121
```

B

```
LMD-9_Cas1_Cr      1 MTWRVVHVSQSEKMRLKLDNLLVQKMGQEFTVPLSDISIIVAEGGDTVVT     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas1_Cr1      1 MTWRVVHVSQSEKMRLKLDNLLVQKMGQEFTVPLSDISIIVAEGGDTVVT     50

LMD-9_Cas1_Cr     51 LRLLSALSKYNIALVVCDNEHLPTGIYHSQNGHFRAYKRLKEQLDWSQKQ    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas1_Cr1     51 LRLLSALSKYNIALVVCDNEHLPTGIYHSQNGHFRAYKRLKEQLDWSQKQ    100

LMD-9_Cas1_Cr    101 KDKAWQIVTYYKINNQEDVLAMFEKSLDNIRLLSDYKEQIEPGDRTNREG    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas1_Cr1    101 KDKAWQIVTYYKINNQEDVLAMFEKSLDNIRLLSDYKEQIEPGDRTNREG    150

LMD-9_Cas1_Cr    151 HAAKVYFNELFGKQFVRVTQKEADVINAGLNYGYAIMRAQMARIVAGYGL    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas1_Cr1    151 HAAKVYFNELFGKQFVRVTQKEADVINAGLNYGYAIMRAQMARIVAGYGL    200

LMD-9_Cas1_Cr    201 NGLLGIFHKNEYNQFNLVDDLMEPFRQIVDVWVYDNLRDQEFLKYEYRLG    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas1_Cr1    201 NGLLGIFHKNEYNQFNLVDDLMEPFRQIVDVWVYDNLRDQEFLKYEYRLG    250

LMD-9_Cas1_Cr    251 LTDLLNAKIKYGKETCSVTVAMDKYVKGFIKYISEKDSSKFHCPVVSSLE    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas1_Cr1    251 LTDLLNAKIKYGKETCSVTVAMDKYVKGFIKYISEKDSSKFHCPVVSSLE    300

LMD-9_Cas1_Cr    301 WRK    303
                     |||
DGCC_Cas1_Cr1    301 WRK    303
```

C

```
LMD-9_Cas2_Cr      1 MRYEALRLLCFFDLPMESKDEKRIYRNFRKELISNGFEMLQFSVYYRTCP     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas2_Cr1      1 MRYEALRLLCFFDLPMESKDEKRIYRNFRKELISNGFEMLQFSVYYRTCP     50

LMD-9_Cas2_Cr     51 NRSFANKFYKKLKMSNLPAGNVRLLAVTEKQFSEMTLIIGGKTKQEEIVS    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Cas2_Cr1     51 NRSFANKFYKKLKMSNLPAGNVRLLAVTEKQFSEMTLIIGGKTKQEEIVS    100

LMD-9_Cas2_Cr    101 DNKLVAI    107
                     |||||.|
DGCC_Cas2_Cr1    101 DNKLVVI    107
```

D

```
LMD-9_Csn2_Cr      1 MKFFVRHPYKDRIELNIGAITQIVGQNNELKYYILQILSWYFGGKKYSRE     50
                     |||||:||||:|||||||||||||||||||..|||:|||||||||.|
DGCC_Csn2_Cr1      1 MKFFVQHPYKERIELNIGAITQIVGQNNELKYYTWQILNWYFGGKKYSSE     50

LMD-9_Csn2_Cr     51 DLSIFDYEEPTILDEAREIVKRSSYHYIDISTFKDLLEQMEYKKGTLAHG    100
                     ||||||||||||||..|||||:||||||||:||||||||||||||||||.
DGCC_Csn2_Cr1     51 DLSIFDYEEPTILDEVGEIVKRNSYHYIDISSFKDLLEQMEYKKGTLAHA    100

LMD-9_Csn2_Cr    101 YLRKIVNQVDIVGHLEKINEQVELIEEAMNRHINLNCGQVEYHLENLPLT    150
                     |||||.||||||.|||||||||||||.|||:||||||||||||||||||||
DGCC_Csn2_Cr1    101 YLRKIANQVDIVTHLEKINEQVELIERAMNQHINLNCGQVEYHLENLPLT    150

LMD-9_Csn2_Cr    151 LDQLLTKNFSPFFTIENKNLSFEWVSNIDKLSLFLEMLDHLLLQTTEKYL    200
                     |||||||||||||.|||||||||||||||||||||||||||||.|||||||
DGCC_Csn2_Cr1    151 LDQLLTKNFSPFFAIENKNLSFEWVSNIDKLSLFLEMLDHLLSQTTEKYL    200

LMD-9_Csn2_Cr    201 IVLKNIDGFISEESYTIFYRQICHLVKKYPNLTFILFPSDQGYLKIDEEN    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Csn2_Cr1    201 IVLKNIDGFISEESYTIFYRQICHLVKKYPNLTFILFPSDQGYLKIDEEN    250

LMD-9_Csn2_Cr    251 SRFVNILSDQVEHLYDVEFMYERVMKYYPSNDFPTREGFRMSLETVTPYL    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
DGCC_Csn2_Cr1    251 SRFVNILSDQVEHLYDVEFMYERVMKYYPSNDFPTREGFRMSLETVTPYL    300

LMD-9_Csn2_Cr    301 LTKMLRQPSLSLVDSVILNILNQLFHFSYRIRYLQKSDEKLLQKFLESKD    350
                     ||||||||||||||||||||||||||||||||.|..|:::||.|||||||
DGCC_Csn2_Cr1    301 LTKMLRQPSLSLVDSVILNILNQLFHFSYRIRYSQTPDKELLHKFLESKD    350
```

Supplementary Figure S2. Sequence similarity of the Cas proteins between *S. thermophilus* LMD-9 and DGCC7710 strains.

The amino acids sequence alignments of **(A)** Cas9, **(B)** Cas1, **(C)** Cas2 and **(D)** Csn2 between *S. thermophilus* LMD-9 and DGCC7710 strains. Pairwise sequence alignment (EMBOSS Needle) was used for analyzing the sequence similarity.

A

```
CRISPR1            1 caagaac---agttat--tgattttataatcactat-gtgggtatga--a     42
                       |   |||.||  |||..|||        || ||||.| ||.|  |
CRISPR3            1 ------ctgaagtcatgctgagatta--------atagtgcg-attacga     35

CRISPR1           43 aatct----caaaa---atcatttgag     62
                     |||||    .||||   |||.|..|||
CRISPR3           36 aatctggtagaaaagatatcctacgag     62
```

B

```
CRISPR1            1 gtttt-------tgtactctcaagat--tt--aagtaactgtacaac     36
                     |||||       |||..|.||.| ||  ||  ||  |||
CRISPR3            1 gttttagagctgtgttgtttcga-atggttccaa--aac--------     36
```

Supplementary Figure S3. Sequence similarity of the leader and repeat sequences between *S. thermophilus* LMD-9 CRISPR1 and CRISPR3 loci.

(A) The pairwise sequence alignments showed that the leader sequences between CRISPR1 and CRISPR3 share only 49.4% of identity. The leader-anchoring sequence (LAS) of the CRISPR3 is highlighted. **(B)** The sequence identity of the repeat sequences between CRISPR1 and CRISPR3 is only 44.7%. Pairwise sequence alignment (EMBOSS Needle) was used for analyzing the sequence similarity.

A

>Cas1 (CRISPR1)

MTWRVVHVSQSEKMRLKLDNLLVQKMGQEFTVPLSDISIIVAEGGDTVVTLRLLSALSKYNIALVVCDNEHL**PT**
GIYHSQNGHFRAYKRLKEQLDWSQKQKDKAWQIVTYYKINNQEDVLAMFEKSLDNIRLLSDYKEQIEPGDR
TNREGHAAKVYFNELFGKQFVRVTQKEADVINAGLNYGYAIMRAQMARIVAGYGLNGLLGIFHKNEYNQFNL
VDDLMEPFRQIVDVWVYDNLRDQEFLKYEYRLGLTDLLNAKIKYGKETCSVTVAMDKYVKGFIKYISEKDSSK
FHCPVVSSLEWRK

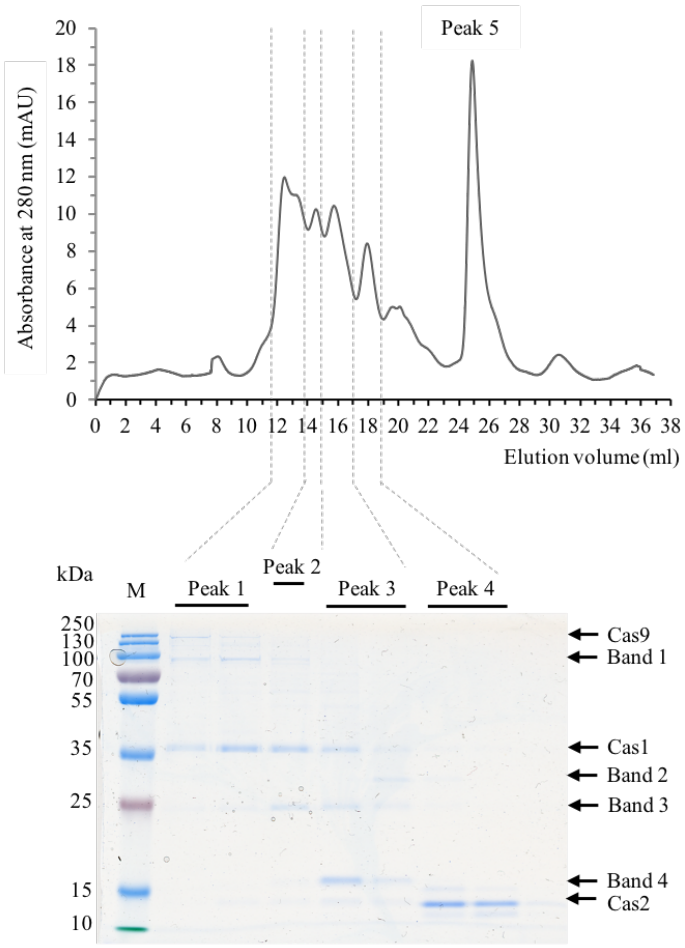
B

>Cas1 (CRISPR1)

MTWRVVHVSQSEKMRLKLDNLLVQKMGQEFTVPLSDISIIVAEGGDTVVTLRLLSALSKYNIALVVCDNEHLPT
GIYHSQNGHFRAYKRLKEQLDWSQKQKDKAWQIVTYYKINNQEDVLAMFEKSLDNIRLLSDYKEQIEPGDRTN
REGHAAKVYFNELFGKQFVRVTQKEADVINAGLNYGYAIMRAQMARIVAGYGLNGLLGIFHKNEYNQFNLVD
DLMEPFRQIVDVWVYDNLRDQEFLKYEYRLGLTDLLNAKIK**YGKETCSVTVAMDKYVKGFIKYISEKDSSKF**
HCPVVSSLEWRK

Supplementary Figure S4. Cas1 SPOT assay reveals the interacting regions.

The amino acids sequences of Cas1 involved **(A)** in Cas1 dimerization (region II; residues 73 to 112; β-sheet-8 to α-helix-3), and **(B)** in the interaction with Cas9 (Cas9 was pre-incubated with tracrRNA) (region VII; residues 261 to 292; β-sheet-10 to α-helix-10) are highlighted and in bold.



Supplementary Figure S5. The study of Cas9-Cas1-Cas2 complex formation via size-exclusion chromatography.

The cell lysates of the individually expressed recombinant N-terminal His_6-tagged Cas9, C-terminal His_6-tagged Cas1 and N-terminal SUMO-His_6-tag of Cas2 were combined and purified via immobilized metal affinity chromatography (IMAC) with TALON Co^{2+} resin. The N-terminal SUMO-His_6-tag of Cas2 was cleaved off by adding the SUMO protease to the protein mixture before applying the proteins mixture to the size-exclusion chromatography (SEC) column Superdex® 200 Increase 10/300 GL (GE Healthcare). The eluates were analyzed on Coomassie-stained 12% SDS-PAGE gel, which showed that Cas9 elutes as peak 1, Cas1 eluates as peak 1, 2 and 3, whereas Cas2 eluates as peak 4. Peak 5 is the elution of imidazole. Cas9, 120.4 kDa; Cas1, 36.2 kDa; Cas2, 12.6 kDa. M, PageRuler™ Plus Pre-stained Protein Ladder (Thermo Scientific).

9 Supplementary Tables

Supplementary Table S1. The tested conditions and the corresponding results for *S. pyogenes* type II-A spacer acquisition assay.

	Heterologous system			***S. pyogenes* endogenous system**	
	Two-plasmid system	**Three-plasmid system**	**Two-plasmid system with phage challenge**		
Strain	*E. coli* BL21-AI	*E. coli* BL21-AI	*E. coli* BL21-AI	*S. pyogenes* WT	*S. pyogenes* Δ*cas9*
Vectors[a]	[1] pCDF-DUETΩ*cas1-cas2-csn2*	[1] pCDF-DUETΩ *cas1-cas2-csn2*	[1] pCDF-DUETΩ *cas1-cas2-csn2*	pEC85	(only transformed with one of the vector below)
	[2] pEC85ΩtracrRNA-Leader-CRISPR	[2] pEC85ΩtracrRNA-Leader-CRISPR-*cas9*	[2] pEC85ΩtracrRNA-Leader-CRISPR-*cas9*		[1] pEC85
		[3] pUC19ΩSpy_0700-NTG (protospacer_mutated-PAM)			[2] pEC85ΩspeM (protospacer)
					[3] pEC85ΩtracrRNA-*cas9*-D10A-speM (a Cas9 RuvC mutant)
					[4] pEC85ΩtracrRNA-*cas9*-H840A-speM (a Cas9 HNH mutant)
Source of new spacers	Plasmids	Plasmids	Phage	Plasmids	Plasmids
Conditions	2 setups: (1) No antibiotics; (2) With Str and Kan during sub-culturing	Only Str and Kan were used (no Amp) during sub-culturing	WT Cas9 was included to allow interference after spacer acquisition	No antibiotics were used during sub-culturing	Either Cas9 was absent or a Cas9 RuvC or HNH mutant was supplemented.
			MOIs of 1, 10 and 100		Kanamycin was used during sub-culturing
Results	No spacer acquisition was detected in *S. pyogenes* and *E. coli* CRISPR arrays	No spacer acquisition was detected in *S. pyogenes* and *E. coli* CRISPR arrays	No spacer acquisition was detected in *S. pyogenes* and *E. coli* CRISPR arrays	No spacer acquisition was detected	No spacer acquisition was detected
	Figure 2B	Figure 2C	Figure 2D	(Figure not shown)	Figure 3

pCDF-DUET, streptomycin resistant (Str); pEC85, kanamycin resistant (Kan); pUC19, ampicillin resistant (Amp). [a] Multiple vectors were used in the bacteria, unless otherwise indicated.

Supplementary Table S2. Sequence similarities of elements of the type II-A CRISPR-Cas systems of *S. pyogenes* SF370 and *S. thermophilus* LMD-9.

Elements	*Sth*_Cr1 vs *Sth*_Cr3		*Sth*_Cr1 vs *Spy*		*Sth*_Cr3 vs *Spy*	
	Identity	Similarity	Identity	Similarity	Identity	Similarity
Cas1	31.6%	48.9%	32.9%	50.0%	79.6%	91.3%
Cas2	37.1%	54.3%	36.0%	55.3%	86.0%	93.0%
Csn2	13.3%	23.4%	15.0%	30.4%	56.8%	80.5%
Cas9	18.9%	32.8%	21.0%	35.0%	57.7%	73.6%
tracrRNA	48.3%	48.3%	51.4%	51.4%	81.2%	81.2%

*Sth*_Cr1, S. *thermophilus* LMD-9, CRISPR1; *Sth*_Cr3, S. *thermophilus* LMD-9, CRISPR3; *Spy*, *S. pyogenes* SF370 (M1 GAS). The sequence identity and similarity were analyzed by pairwise sequence alignment (EMBOSS Needle). Amino acids sequence identity and similarity were demonstrated for Cas proteins, whereas nucleotide sequences identity and similarity were shown for tracrRNA.

Supplementary Table S3. Sequence similarities of the Cas proteins between *S. thermophilus* LMD-9 and DGCC7710 strains (CRISPR1)

Protein	Identity	Similarity[a]
Cas1	100.0%	100.0%
Cas2	99.1%	99.1%
Csn2	93.4%	95.7%
Cas9	99.8%	99.9%

The sequence similarities were analyzed using Pairwise Sequence Alignment (EMBOSS Needle).

Supplementary Table S4. Yeast two-hybrid analysis of the interacting partners of Cas1 (Spy_1047) of *S. pyogenes* SF370 and their corresponding *S. thermophilus* LMD-9 orthologs.

PBS[1]	Interacting partners of Cas1 (*S. pyogenes* SF370)	Functions	Orthologs for *S. thermophilus* LMD-9
A	**Cas1** (Spy_1047)	• CRISPR adaptation	**Cas1** (CRISPR3) (STER_1476) • Identity: 0.796
C	CMP-binding factor, **Cbf** (Spy_0267)	• Metal ion binding • Nucleic acid binding • Phosphoric diester hydrolase activity	CMP-binding protein (STER_1766) • Identity: 0.848
C	**Csn2** (Spy_1049)	• Binds dsDNA	**Csn2** (CRISPR3) (STER_1474) • Identity: 0.573
C	Ribonuclease HII, **RnhB** (Spy_1162)	• Endonuclease that specifically degrades the RNA of RNA-DNA hybrids • DNA replication	Ribonuclease HII, **RnhB** (STER_0920) • Identity: 0.676

[1] Predicted Biological Score (PBS) shows the reliability of the interaction, *i.e.* PBS-A indicates a very high reliability in the interaction, B indicates high reliability, C indicates good reliability. The candidates with PBS-D were not analyzed in this study (except for Csn2 analysis), as they show moderate reliability in the interaction. This means that PBS-D includes a mix false-positive or hardly detectable interaction, and they need to be carefully verified. Due to the low reliability in the interaction for the candidates with PBS-E and F, these candidates were not analyzed here.

Supplementary Table S5. Yeast two-hybrid analysis of the interacting partners of Cas2 (Spy_1048) of *S. pyogenes* SF370 and their corresponding *S. thermophilus* LMD-9 orthologs.

PBS[1]	Interacting partners of Cas2 (*S. pyogenes* SF370)	Functions	Orthologs for *S. thermophilus* LMD-9
A	Cell-cycle protein (Spy_0013)	• tRNA modification	Cell-cycle protein, **MesJ/Ycf62** family (STER_0012) • Identity: 0.501
A	Tyrosyl-tRNA synthetase, **TyrS** (Spy_0096)	• Aminoacyl-tRNA biosynthesis	Tyrosyl-tRNA synthetase (STER_1847) • Identity: 0.837
A	RNA methyltransferase (Spy_1346)	• RNA binding • RNA processing	tRNA methyltransferase, **TrmA** family (STER_0702) • Identity: 0.776
A	Exodeoxyribonuclease VII large subunit, **XseA** (Spy_1500)	• Mismatch repair	Exodeoxyribonuclease VII large subunit, **XseA** (STER_1184) • Identity: 0.742
A	Excinuclease ABC subunit A, **UvrA** (Spy_1825)	• Nucleotide excision repair	Excinuclease ABC subunit A, **UvrA** (STER_1722) • Identity: 0.876
B	**Cas1** (Spy_1047)	• CRISPR adaptation	**Cas1** (CRISPR3) (STER_1476) • Identity: 0.796
B	GTP-binding protein **TypA** (Spy_1527)	• GTPase activity	GTP-binding protein **TypA/BipA** (STER_0771) identity: 0.941
C	Cell envelope proteinase, **PrtS** (Spy_0416)	• Serine-type endopeptidase activity	Subtilisin-like serine protease (STER_0846) • Identity: 0.427
C	30S ribosomal protein S1, **RpsA** (Spy_0913)	• DNA binding • Structural constituent of ribosome	30S ribosomal protein S1 (STER_0639) • Identity: 0.885
C	trigger factor, **Tig** (Spy_1896)	• Promoting folding of newly synthesized proteins	Trigger factor (STER_0191) • Identity: 0.775

[1] Predicted Biological Score (PBS) shows the reliability of the interaction, *i.e.* PBS-A indicates a very high reliability in the interaction, B indicates high reliability, C indicates good reliability. The candidates with PBS-D were not analyzed in this study (except for Csn2 analysis), as they show moderate reliability in the interaction. This means that PBS-D includes a mix false-positive or hardly detectable interaction, and they need to be carefully verified. Due to the low reliability in the interaction for the candidates with PBS-E and F, these candidates were not analyzed here.

Supplementary Table S6. Yeast two-hybrid analysis of the interacting partners of Csn2 (Spy_1049) of *S. pyogenes* SF370 and their corresponding *S. thermophilus* LMD-9 orthologs.

PBS[1]	Interacting partners of Csn2 (*S. pyogenes* SF370)	Functions	Orthologs for *S. thermophilus* LMD-9
A	Queuine tRNA-ribosyltransferase, **Tgt** (SPy_0203)	• Queuosine biosynthetic process	Queuine tRNA-ribosyltransferase (STER_1785) • Identity: 0.903
A	NAD-dependent DNA ligase **LigA** (SPy_0751)	• Nucleotide excision repair • Base excision repair • Mismatch repair • DNA replication	NAD-dependent DNA ligase **LigA** (STER_1513) • Identity: 0.730
A	30S ribosomal protein S7, **RpsG** (SPy_0272)	• Translation	30S ribosomal protein S7 (STER_1763) • Identity: 0.955
A	Serine hydroxymethyltransferase, **GlyA** (Spy_1145)	• Glycine biosynthetic process from serine • Pyridoxal phosphate binding	Serine hydroxymethyltransferase (STER_0796) • Identity: 0.755
B	N-acetylmuramoyl-L-alanine amidase (Spy_1764)	• Amidase activity	Peptidoglycan hydrolase (STER_0160) Identity: 0.361
B	tRNA/rRNA methyltransferase (Spy_1938)	• RNA binding • RNA processing	tRNA/rRNA methyltransferase (STER_0122) • Identity: 0.84
C	16S rRNA methyltransferase **GidB** (Spy_0329)	• RNA methytransferase activity	16S rRNA methyltransferase GidB (STER_0352) • Identity: 0.751
C	Aminodeoxychorismate lyase (Spy_0348)	• Lyase activity	Aminodeoxychorismate lyase (STER_0288) • Identity: 0.459
C	Branched-chain alpha-keto acid dehydrogenase E2 subunit, **AcoC** (Spy_1029)	• Dihydrolipoyllysine-residue acetyltransferase activity	Branched-chain alpha-keto acid dehydrogen (STER_1034) • Identity: 0.764
C	Penicillin-binding protein 2a, **Pbp2A** (Spy_2059)	• Penicillin binding • Transferase activity	Penicillin-binding protein 2a (STER_0260) • Identity: 0.632
D	**Cas9** (SPy_1046)	• CRISPR interference • CRISPR adaptation	**Cas9** (CRISPR3) (STER_1477) • Identity: 0.578

[1] Predicted Biological Score (PBS) shows the reliability of the interaction, *i.e.* PBS-A indicates a very high reliability in the interaction, B indicates high reliability, C indicates good reliability. The candidates with PBS-D were not analyzed in this study (except for Csn2 analysis), as they show moderate reliability in the interaction. This means that PBS-D includes a mix false-positive or hardly detectable interaction, and they need to be carefully verified. Due to the low reliability in the interaction for the candidates with PBS-E and F, these candidates were not analyzed here.

Supplementary Table S7. Yeast two-hybrid analysis of the interacting partners of Cas9 (Spy_1046) of *S. pyogenes* SF370 and their corresponding *S. thermophilus* LMD-9 orthologs.

PBS[1]	Interacting partners of Cas9 (*S. pyogenes* SF370)	Functions	Orthologs for *S. thermophilus* LMD-9
A	Excinuclease ABC subunit A, **UvrA** (Spy_1825)	• Nucleotide excision repair	Excinuclease ABC subunit A, **UvrA** (STER_1722) • Identity: 0.876
B	DNA helicase II / ATP-dependent DNA helicase **PcrA** (SPy_1267)	• Nucleotide excision repair • Mismatch repair	ATP-dependent DNA helicase PcrA (STER_0994) • Identity: 0.820

[1] Predicted Biological Score (PBS) shows the reliability of the interaction, *i.e.* PBS-A indicates a very high reliability in the interaction, B indicates high reliability, C indicates good reliability. The candidates with PBS-D were not analyzed in this study (except for Csn2 analysis), as they show moderate reliability in the interaction. This means that PBS-D includes a mix false-positive or hardly detectable interaction, and they need to be carefully verified. Due to the low reliability in the interaction for the candidates with PBS-E and F, these candidates were not analyzed here.

Supplementary Table S8. The interacting partners of Cas1 identified by *in vitro* pull-down assay in combination with mass spectrometry.

System/ Pathway	Identified Proteins	Accession Number	KEGG ID	Fold Change[a]
CRISPR	CRISPR-associated endonuclease, **Cas9** family	gi|116101490	STER_0709	8.8
	CRISPR-associated protein, **Cas1** family	gi|116100820	STER_0710	4.2
DNA Repair	DNA mismatch repair protein **MutS**	gi|116100292	STER_0068	3.6
	Excinuclease ABC subunit B, **UvrB**	gi|116101472	STER_1457	3.5
	DNA helicase/exodeoxyribonuclease V, subunit B, **AddB**	gi|116101679	STER_1682	3.4
	DNA helicase/exodeoxyribonuclease V, subunit A, **AddA**	gi|116101678	STER_1681	3.1
	Exonuclease **RecJ**	gi|116101235	STER_1191	3.1
ATP Binding Cassette (ABC) Transporters	ATPase component of ABC transporter with duplicated ATPase domains	gi|116101279	STER_1237	3.7
	ABC-type polysaccharide/polyol phosphate transport system, ATPase component	gi|116101450	STER_1434	3.2
	ATPase component of ABC transporter with duplicated ATPase domains	gi|116100599	STER_0462	3.2
	ABC-type polar amino acid transport system, ATPase component	gi|116101618	STER_1617	3.0
DNA Replication	Ribonucleoside-triphosphate reductase class III catalytic subunit / ribonucleoside-triphosphate reductase	gi|116101906	STER_1942	4.0
	DNA polymerase III catalytic subunit, STER_1935 type	gi|116100303	STER_0095	3.5
	DNA primase	gi|116101465	STER_1449	3.2

[a] Only candidates with fold change ≥3.0 are shown.

Supplementary Table S9. Buffers used for protein purification.

Experiment[1]	*in vitro* pull-down assay and SPOT peptide assay	*in vitro* pull-down assay	Protein-protein interaction study: SEC[2] and crosslinking assays
Protein	CPD-His_{12}-tagged Cas1 and His6-tagged Cas1	His_6-tagged Cas9	His_6-tagged Cas1, SUMO-His_6-tag Cas2 and His_6-tagged Cas9
Lysis buffer	10 mM Tris-HCl, pH8.0; 300 mM NaCl; 2.5 mM β-Mercaptoethanol; 10% glycerol	20 mM HEPES-KOH, pH7.5; 1 M KCl; 0.1% Triton X-100; 25 mM imidazole	20 mM HEPES-HCl, pH 6.8; 500 mM NaCl; 10% glycerol; 10 mM imidazole; 2 mM TCEP (Sigma); 0.1% Triton X-100 (Serva Electrophoresis); 0.5 mM PMSF (Sigma-Aldrich); Complete EDTA-free protease inhibitor (Roche)
Wash buffer	(same as lysis buffer)	(same as lysis buffer)	20 mM HEPES-HCl, pH 6.8; 1 M NaCl; 10% glycerol; 25 mM imidazole; 2 mM TCEP
Elution buffer	10 mM Tris-HCl, pH8.0; 300 mM NaCl; 2.5 mM β-Mercaptoethanol; 10% glycerol; 500 mM imidazole	20 mM HEPES-KOH, pH7.5; 150 mM KCl; 0.1 mM DTT; 250 mM imidazole; 1 mM EDTA	20 mM HEPES-HCl, pH 6.8; 250 mM NaCl; 10% glycerol; 0.05% Tween-20; 2 mM TCEP; a gradient of imidazole, 150, 250 and 500 mM
Buffer A	(N/A)	20 mM HEPES-KOH, pH7.5; 100 mM KCl	(only for Cas9) 20 mM HEPES, pH 6.8; 250 mM NaCl; 0.05% Tween-20; 10% glycerol; 2 mM TCEP
SEC buffer	10 mM Tris-HCl, pH8.0; 300 mM NaCl; 2.5 mM β-Mercaptoethanol; 10% glycerol	(N/A)	20 mM HEPES-HCl, pH 6.8; 250 mM NaCl; 10% glycerol; 0.05% Tween-20

[1] The specific assay that involved the purified protein

[2] Size-exclusion chromatography is abbreviated as SEC

N/A indicates not applicable

Supplementary Table S10. List of bacterial and viral strains

Strain	Relevant characteristics	Source
Streptococcus pyogenes		
EC904	SF370 (M1 serotype), WT	ATCC 700294
EC1788	EC904Δ*cas9*	Deltcheva et al, 2011
Streptococcus thermophilus		
EC2162	LMD-9 (WT)	Sylvain Moineau
EC2724	EC2161ΔCR2_ΔCR3	This study
EC2735	EC2724Δ*cas9*_CR1	This study
Escherichia coli		
RDN204	TOP10; Host for cloning	Invitrogen
RDN226	DH5α; Host for cloning	Lab collection
EC1265	BL21 (DE3) Rosetta; Expression expression	Novagen
EC2159	BL21-AI	Stan Brouns
EC2212	NiCo21 (DE3); Expression strain	NEB
Phage for *Streptococcus thermophilus* LMD-9		
Phage DT1		Félix d'Hérelle Reference Center for bacterial viruses of the Université Laval
Phage for *Escherichia coli*		
Phage Lamda (virulent)		Stan Brouns

Supplementary Table S11. List of plasmids

Plasmids	Relevant characteristics	Source
Vectors for *S. pyogenes*		
pEC85	*repDEG*-pAMβ1, pJH1-*aphIII*, ColE1	Lab collection
Vectors for *E. coli*		
pEC180	pET21a	Lab collection
pEC225	pET16b	Lab collection
pEC574	pCDFDUET-1	Lab collection
pEC707	pUC19	Lab collection
pEC1075	pET20b	Lab collection
pEC1076	pEC-A_Hi_SUMO	Lab collection
pEC1078	pEC21-CPD_Pto	Lab collection
Plasmids for *S. pyogenes* (SF370) adaptation study – Heterologous system in *E. coli* BL21-AI		
pEC645	pEC85Ω171tracrRNA-Leader-CRISPR (171 nt form)	This study
pEC646	pEC85Ω89tracrRNA-Leader-CRISPR (89 nt form)	This study
pEC663	pEC645Ω*Pcas9(Spy)-cas9(Spy)*	Lab collection
pEC651	pEC574Ω*cas1-cas2-csn2*	This study
pEC686	pUC19Ωspy0700-GG(PAM)	This study
pEC687	pUC19Ωspy0700-TG(PAM)	This study
Plasmids for endogenous spacer acquisition study in *S. pyogenes* SF370		
pEC488	pEC85ΩspeM (protospacer)	Lab collection
pEC659	pEC85ΩtracrRNA-*cas9*-D10A-speM (a Cas9 RuvC mutant)	Lab collection
pEC660	pEC85ΩtracrRNA-*cas9*-H840A-speM (a Cas9 HNH mutant)	Lab collection
Plasmids for *S. thermophilus* (LMD-9) adaptation study – Heterologous system in *E. coli* BL21-AI		
pEC1151	pEC574Ω*cas1-cas2-csn2(Sth_Cr1)*	This study
pEC1230	pEC85ΩCR1_5spcs_Sth	This study
Plasmids for over-expression of *S. thermophilus* (LMD-9) Cas proteins		
pEC621	pEC225 inserted with cassette harboring NotI, SacI, SalI site	Fonfara et al, 2013
pEC641	pEC621Ω*cas9*(Sth_Cr1)	Fonfara et al, 2013
pEC1290	pEC1075Ω*cas1(Sth_Cr1)*	This study
pEC1295	pEC1076Ω*cas2*(Sth_Cr1)	This study
pEC1300	pEC1078Ω*cas1*(Sth_Cr1)	This study
Plasmids for endogenous spacer acquisition study in *S. thermophilus* (LMD-9) – CRISPR1 locus		
pEC2376	pWAR2228Ω*cas1-cas2-csn2-cas9*_Cr1(Sth_DGCC7710)	Wei et al, 2015
pEC2377	pWAR2228Ω*cas1-cas2-csn2*_Cr1(Sth_DGCC7710)	Wei et al, 2015

Spy, *Streptococcus pyogenes* SF370; Sth, *Streptococcus thermophilus* LMD-9; Sth_DGCC7710, *Streptococcus thermophilus* DGCC7710

Supplementary Table S12. List of primers

Purpose	Primer code	Sequence 5'-3'[a]	F/R[b]	Usage
Plasmids for *S. pyogenes* (SF370) adaptation study – Heterologous system in *E. coli* BL21-AI				
pEC645	OLEC2968	GCATCGGGATCC*GTTTGCAGTCAGAGTAGAAT AGAAGTATC*	F	Cloning 171-tracrRNA
	OLEC2969	GCGGAAAATCATATAGTTCACGTGGCCTGC AGG*TTAATGACCTCCGAAATTAGTTTAATATGC*	R	
	OLEC2971	GCATATTAAACTAATTTCGGAGGTCATTAAC CTGCAGG*CCACGTGAACTATATGATTTTCCGC*	F	Cloning Leader-CRISPR
	OLEC2967	GGTGGTGAATTC*GCTTTAACAGAAAGAATAGG AAGGTATCC*	R	
	OLEC2968	GCATCGGGATCC*GTTTGCAGTCAGAGTAGAAT AGAAGTATC*	F	LM-PCR; Cloning 171-tracrRNA-Leader-CRISPR
	OLEC2967	GGTGGTGAATTC*GCTTTAACAGAAAGAATAGG AAGGTATCC*	R	
pEC646	OLEC2968	GCATCGGGATCC*GTTTGCAGTCAGAGTAGAAT AGAAGTATC*	F	Cloning 89-tracrRNA
	OLEC2970	GCGGAAAATCATATAGTTCACGTGGCCTGC AGG*GAATTTCTCCTTGATTATTTGTTATAAAAG*	R	
	OLEC2972	CTTTTATAACAAATAATCAAGGAGAAATTCC CTGCAGG*CCACGTGAACTATATGATTTTCCGC*	F	Cloning Leader-CRISPR
	OLEC2967	GGTGGTGAATTC*GCTTTAACAGAAAGAATAGG AAGGTATCC*	R	
	OLEC2968	GCATCGGGATCC*GTTTGCAGTCAGAGTAGAAT AGAAGTATC*	F	LM-PCR; Cloning 89-tracrRNA -Leader-CRISPR
	OLEC2967	GGTGGTGAATTC*GCTTTAACAGAAAGAATAGG AAGGTATCC*	R	
pEC651	OLEC2973	ATGCAGGGATCC*ATGGCTGGTTGGCGTACTG TTGTG*	F	Cloning *cas1cas2csn2*
	OLEC1757	CTGCATGAATTC*TTATACCATATTTTTAGTTA*	R	
pEC663	OLEC3040	ATGACTCTAGA*GGAGAAATTCAAAGAAATTTAT CAGC*	F	Cloning ΩP*cas9(Spy)-cas9(Spy)*
	OLEC3041	ATGACTCTAGA*AACCAGCCATCAGTCACCTC*	R	
pEC686	OLEC2106	ATGCAGGGCCGGCC*AGTATCAGCGTACTTGG ATTTGGG*	F	Cloning spy0700-GG(PAM); Source pEC573
	OLEC2107	ATGCAGGGCCGGCC*TTGTCTCACTCACTCTAT TTTTG*	R	
pEC687	OLEC2106	ATGCAGGGCCGGCC*AGTATCAGCGTACTTGG ATTTGGG*	F	Cloning
	OLEC2107	ATGCAGGGCCGGCC*TTGTCTCACTCACTCTAT TTTTG*	R	
RT-PCR	OLEC3048	*GTTACCAATATGAGGAAGATTCTGAAC*	F	Checking the expression of *csn2*
	OLEC3049	*CATTCAAAGACAATCAGTTCTGTAATC*	R	
Spacer acquisition screening	OLEC1749	GGTGGTGGATCC*CCACGTGAACTATATGATTT TCCGC*	F	For the type II-A array of *S. pyogenes* (primer set 1)
	OLEC3127	*CGCAAGAAGAAATCAACCAGCG*	R	
	OLEC1749	GGTGGTGGATCC*CCACGTGAACTATATGATTT TCCGC*	F	For the type II-A array of *S. pyogenes* (primer set 2)
	OLEC2967	GGTGGTGAATTCGCTTTAACAGAAAGAATA GGAAGGTATCC	R	
	OLEC3066	*GACTGAAGTCTAGCTGAGAC*	F	For the type II-A array of *S. pyogenes* (primer set 3)
	OLEC1141	*CAAATTGAGTTATGTTCATATAAG*	R	
	OLEC3005	GGTGGTGGATCC*GCGGTAAAGTTTGGTAGATT TTAGTTTG*	F	For the CRISPR I array of *E. coli* BL21-AI
	OLEC3006	GGTGGTCTGCAG*GTTACGTGGATATGTTGCTT ATTACAAG*	R	
	OLEC3007	GGTGGTGGATCC*CCAGTTATCGTGAGAGTAAT TCATCG*	F	For the CRISPR II array of *E. coli* BL21-AI
	OLEC3008	GGTGGTCTGCAG*CGTGATGTTATGCGGATAAT GCTACC*	R	

Purpose	Primer code	Sequence 5'-3'[a]	F/R[b]	Usage
Plasmids for Cas1 over-expression (*S. thermophilus* LMD-9; CRISPR1)				
pEC1290	OLEC4721	GGAGAACATATG*ACTTGGAGAGTTGTACATG*	F	Cloning
	OLEC4722	TACCTCCTCGAG*TTTTCTCCACTCTAAACTTG*	R	
pEC1300	OLEC4721	GGAGAACATATG*ACTTGGAGAGTTGTACATG*	F	Cloning
	OLEC4722	TACCTCCTCGAG*TTTTCTCCACTCTAAACTTG*	R	
Plasmids for Cas2 over-expression (*S. thermophilus* LMD-9; CRISPR1)				
pEC1295	OLEC4495	*ACCAGGAACAAACCGGCGGCCGCTCGATGAG GTATGAAG*	F	Cloning
	OLEC4496	*GCAAAGCACCGGCCTCGTTATATGGCCACCAA C*	R	
Plasmids for *S. thermophilus* (LMD-9) adaptation study – Heterologous system in *E. coli* BL21-AI				
pEC1151	OLEC4518	GGTGGTCCATGGGC*ACTTGGAGAGTTGTACAT GTCAGTC*	F	Cloning
	OLEC4523	GGTGGTGGATCC*TCAATCCTTACTTTCTAAAAA TTTC*	R	
pEC1230	OLEC 4833	CCGCTGCATGCCTGCAGGTCGACTCTAGAG GATCCGTTACCGTATAAGATATTTACAAAAA TC	F	Cloning
	OLEC 4841	CTTTTTATGTTGAATCAA*CTATTTCACGATATT TTTTCACGAAT*	R	
	OLEC 4842	ATTCGTGAAAAAATATCGTGAAATAG*TTGAT TCAACATAAAAAGCCGG*	F	
	OLEC 4838	CACTTTTGTGGGCCTTTTTTGGCCGGCCTGA ATTC*GAGCCTCCCTATCCTTAATTTG*	R	
	OLEC 4839	*GAATTCAGGCCGGCCAAAAAAGG*	F	
	OLEC 4840	*GGATCCTCTAGAGTCGACCTGCAG*	R	
RT-PCR	OLEC4522	GGTGGTCCATGGGC*AAATTTTTTGTACGACAT CCTTAC*	F	Checking the expression of *csn2*
	OLEC4523	GGTGGTGGATCC*TCAATCCTTACTTTCTAAAAA TTTC*	R	
Spacer acquisition screening	OLEC4778	*GATTTTATAATCACTATGTGGG*	F	For the CRISPR1 array
	OLEC4779	*GATGGTCGGTTATTTTTCAG*	R	
	OLEC4780	*GGTGACAGTCACATCTTGTC*	F	For the CRISPR3 array
	OLEC4781	*GTTTTCGTCTTGGATACCAC*	R	

[a] *italic*, sequence annealing to the template; underlined, restriction site; No formating, extra nucleotides.
[b] F, forward primer; R, reverse primer.

10 References

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569–573.

Andersson, A.F., and Banfield, J.F. (2008). Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. Science *320*, 1047–1050.

Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H., et al. (2013). Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. Nucleic Acids Res *41*, 6347–6359.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., et al. (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol *79*, 484–502.

Barrangou, R., and van der Oost, J. (2013). CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea (Springer).

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Barrangou, R., Coûté-Monvoisin, A.-C., Stahl, B., Chavichvily, I., Damange, F., Romero, D.A., Boyaval, P., Fremaux, C., and Horvath, P. (2013). Genomic impact of CRISPR immunization against bacteriophages. Biochem. Soc. Trans. *41*, 1383–1391.

Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., et al. (2008). A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. J. Biol. Chem. *283*, 20361–20371.

Bernheim, A., Calvo-Villamañán, A., Basier, C., Cui, L., Rocha, E.P.C., Touchon, M., and Bikard, D. (2017). Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. Nat. Commun. *8*.

Bhaya, D., Davison, M., and Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet *45*, 273–297.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Kunne, T., Sobota, M., Dekker, C., Brouns, S.J.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol Cell *58*, 60–70.

Bobay, L.M., Touchon, M., and Rocha, E.P. (2013). Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. PLoS Genet *9*, e1003825.

Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D., et al. (2004). Complete sequence and comparative

genome analysis of the dairy bacterium Streptococcus thermophilus. Nat. Biotechnol. *22*, 1554–1558.

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M.F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K.L., and Davidson, A.R. (2015). Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. Nature *526*, 136–139.

Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast Two-Hybrid, a Powerful Tool for Systems Biology. Int. J. Mol. Sci. *10*, 2763–2788.

Brüssow, H. (2001). Phages of Dairy Bacteria. Annu. Rev. Microbiol. *55*, 283–303.

Brussow, H., Canchaya, C., and Hardt, W.-D. (2004). Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. Microbiol. Mol. Biol. Rev. *68*, 560–602.

Caparon, M.G., and Scott, J.R. (1991). Genetic manipulation of pathogenic streptococci. Methods Enzym. *204*, 556–586.

Carapetis, J.R., Beaton, A., Cunningham, M.W., Guilherme, L., Karthikeyan, G., Mayosi, B.M., Sable, C., Steer, A., Wilson, N., Wyber, R., et al. (2016). Acute rheumatic fever and rheumatic heart disease. Nat. Rev. Dis. Primer 15084.

Catalano, C.E., Cue, D., and Feiss, M. (1995). Virus DNA packaging: the strategy used by phage lambda. Mol. Microbiol. *16*, 1075–1086.

Chowdhury, S., Carter, J., Rollins, M.F., Golden, S.M., Jackson, R.N., Hoffmann, C., Nosaka, L., Bondy-Denomy, J., Maxwell, K.L., Davidson, A.R., et al. (2017). Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. Cell *169*, 47-57.e11.

Chylinski, K., Le Rhun, A., and Charpentier, E. (2013). The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. RNA Biol *10*, 726–737.

Chylinski, K., Makarova, K.S., Charpentier, E., and Koonin, E.V. (2014). Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res. *42*, 6091–6105.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun *3*, 945.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol *190*, 1390–1400.

Dillingham, M.S., and Kowalczykowski, S.C. (2008). RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev *72*, 642–671, Table of Contents.

Dong, D., Guo, M., Wang, S., Zhu, Y., Wang, S., Xiong, Z., Yang, J., Xu, Z., and Huang, Z. (2017). Structural basis of CRISPR–SpyCas9 inhibition by an anti-CRISPR protein. Nature *546*, 436–439.

Drabavicius, G., Sinkunas, T., Silanskas, A., Gasiunas, G., Venclovas, Č., and Siksnys, V. (2018). DnaQ exonuclease-like domain of Cas2 promotes spacer integration in a type I-E CRISPR-Cas system. EMBO Rep. *19*, e45543.

Edgar, R.C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Edgar, R.C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Ellinger, P., Arslan, Z., Wurm, R., Tschapek, B., MacKenzie, C., Pfeffer, K., Panjikar, S., Wagner, R., Schmitt, L., Gohlke, H., et al. (2012). The crystal structure of the CRISPR-associated protein Csn2 from Streptococcus agalactiae. J Struct Biol *178*, 350–362.

Elmore, J.R., Sheppard, N.F., Ramia, N., Deighan, T., Li, H., Terns, R.M., and Terns, M.P. (2016). Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR–Cas system. Genes Dev. *30*, 447–459.

Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., Krause, K.L., et al. (2017). Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. Proc Natl Acad Sci U A *114*, E5122–E5128.

Feldmann, E., Schmiemann, V., Goedecke, W., Reichenberger, S., and Pfeiffer, P. (2000). DNA double-strand break repair in cell-free extracts from Ku80-deficient cells: implications for Ku serving as an alignment factor in non-homologous DNA end joining. Nucleic Acids Res *28*, 2585–2596.

Fineran, P.C., Gerritzen, M.J., Suarez-Diez, M., Kunne, T., Boekhorst, J., van Hijum, S.A., Staals, R.H., and Brouns, S.J. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proc Natl Acad Sci U A *111*, E1629-38.

Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lécrivain, A.-L., Bzdrenga, J., Koonin, E.V., and Charpentier, E. (2014). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res. *42*, 2577–2590.

Gardan, R., Besset, C., Guillot, A., Gitton, C., and Monnet, V. (2009). The Oligopeptide Transport System Is Essential for the Development of Natural Competence in Streptococcus thermophilus Strain LMD-9. J. Bacteriol. *191*, 4647–4655.

Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. *109*, E2579–E2586.

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., and Sorek, R. (2015). BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. *34*, 169–183.

Goren, M.G., Doron, S., Globus, R., Amitai, G., Sorek, R., and Qimron, U. (2016). Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell Rep *16*, 2811–2818.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics *8*, 172.

Hare, J.M., Ferrell, J.C., Witkowski, T.A., and Grice, A.N. (2014). Prophage Induction and Differential RecA and UmuDAb Transcriptome Regulation in the DNA Damage Responses of Acinetobacter baumannii and Acinetobacter baylyi. PLoS One *9*, e93861.

Hegge, J.W., Swarts, D.C., and van der Oost, J. (2017). Prokaryotic Argonaute proteins: novel genome-editing tools? Nat. Rev. Microbiol. *16*, 5–11.

Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D., and Marraffini, L.A. (2015). Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature *519*, 199–202.

Hille, F., Richter, H., Wong, S.P., Bratovič, M., Ressel, S., and Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. Cell *172*, 1239–1259.

Hirano, H., Gootenberg, J.S., Horii, T., Abudayyeh, O.O., Kimura, M., Hsu, P.D., Nakane, T., Ishitani, R., Hatada, I., Zhang, F., et al. (2016). Structure and Engineering of Francisella novicida Cas9. Cell *164*, 950–961.

Hochstrasser, M.L., Taylor, D.W., Bhat, P., Guegler, C.K., Sternberg, S.H., Nogales, E., and Doudna, J.A. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc. Natl. Acad. Sci. *111*, 6618–6623.

Hols, P., Hancy, F., Fontaine, L., Grossiord, B., Prozzi, D., Leblondbourget, N., Decaris, B., Bolotin, A., Delorme, C., and Duskoehrlich, S. (2005). New insights in the molecular biology and physiology of revealed by comparative genomics. FEMS Microbiol. Rev. *29*, 435–463.

Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol *190*, 1401–1412.

van Houte, S., Ekroth, A.K.E., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A., et al. (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. Nature *532*, 385–388.

Hudaiberdiev, S., Shmakov, S., Wolf, Y.I., Terns, M.P., Makarova, K.S., and Koonin, E.V. (2017). Phylogenomics of Cas4 family nucleases. BMC Evol. Biol. *17*.

Hyman, P., and Abedon, S.T. (2010). Bacteriophage Host Range and Bacterial Resistance. In Advances in Applied Microbiology, (Elsevier), pp. 217–248.

Hynes, A.P., Rousseau, G.M., Lemay, M.-L., Horvath, P., Romero, D.A., Fremaux, C., and Moineau, S. (2017). An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. Nat. Microbiol. *2*, 1374–1380.

Hynes, A.P., Rousseau, G.M., Agudelo, D., Goulet, A., Amigues, B., Loehr, J., Romero, D.A., Fremaux, C., Horvath, P., Doyon, Y., et al. (2018). Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat. Commun. *9*.

Ivancic-Bace, I., Cass, S.D., Wearne, S.J., and Bolt, E.L. (2015). Different genome stability proteins underpin primed and naive adaptation in E. coli CRISPR-Cas immunity. Nucleic Acids Res *43*, 10821–10830.

Jansen, R., Embden, J.D., Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol *43*, 1565–1575.

Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. Science *348*, 1477–1481.

Jiang, F., Taylor, D.W., Chen, J.S., Kornfeld, J.E., Zhou, K., Thompson, A.J., Nogales, E., and Doudna, J.A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science *351*, 867–871.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., et al. (2014). Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. Science *343*, 1247997–1247997.

Ka, D., Lee, H., Jung, Y.D., Kim, K., Seok, C., Suh, N., and Bae, E. (2016). Crystal Structure of Streptococcus pyogenes Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. Structure *24*, 70–79.

Ka, D., Jang, D.M., Han, B.W., and Bae, E. (2018). Molecular organization of the type II-A CRISPR adaptation module and its interaction with Cas9 via Csn2. Nucleic Acids Res.

Keen, E.C. (2012). Paradigms of pathogenesis: targeting the mobile genetic elements of disease. Front. Cell. Infect. Microbiol. *2*.

Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep. *22*, 3377–3384.

Kim, T.Y., Shin, M., Huynh Thi Yen, L., and Kim, J.S. (2013). Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. Biochem Biophys Res Commun *441*, 720–725.

Klaiman, D., Steinfels-Kohn, E., and Kaufmann, G. (2014). A DNA break inducer activates the anticodon nuclease RloC and the adaptive immunity in Acinetobacter baylyi ADP1. Nucleic Acids Res *42*, 328–339.

Kligler, B., and Cohrssen, A. (2008). Probiotics. Am. Fam. Physician *78*, 1073–1078.

Koo, Y., Jung, D.K., and Bae, E. (2012). Crystal structure of Streptococcus pyogenes Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. PLoS One *7*, e33401.

Koonin, E.V., Makarova, K.S., and Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. *37*, 67–78.

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol. Cell *63*, 852–864.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nat Rev Microbiol *8*, 317–327.

Lee, H., Zhou, Y., Taylor, D.W., and Sashital, D.G. (2018). Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Mol. Cell *70*, 48-59.e5.

Lee, K.H., Lee, S.G., Eun Lee, K., Jeon, H., Robinson, H., and Oh, B.H. (2012). Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. Proteins *80*, 2573–2582.

Lemak, S., Beloglazova, N., Nocek, B., Skarina, T., Flick, R., Brown, G., Popovic, A., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2013). Toroidal Structure and DNA Cleavage by the CRISPR-Associated [4Fe-4S] Cluster Containing Cas4 Nuclease SSO0001 from *Sulfolobus solfataricus*. J. Am. Chem. Soc. *135*, 17476–17487.

Lemak, S., Nocek, B., Beloglazova, N., Skarina, T., Flick, R., Brown, G., Joachimiak, A., Savchenko, A., and Yakunin, A.F. (2014). The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. Nucleic Acids Res. *42*, 11144–11155.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

Li, M., Wang, R., Zhao, D., and Xiang, H. (2014). Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res *42*, 2483–2492.

Liu, T., Liu, Z., Ye, Q., Pan, S., Wang, X., Li, Y., Peng, W., Liang, Y., She, Q., and Peng, N. (2017). Coupling transcriptional activation of CRISPR–Cas system and DNA repair genes by Csa3a in Sulfolobus islandicus. Nucleic Acids Res. *45*, 8978–8992.

Makarova, K.S., Aravind, L., Grishin, N.V., Rogozin, I.B., and Koonin, E.V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res *30*, 482–496.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol *9*, 467–477.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2013). The basic building blocks and evolution of CRISPR–Cas systems. Biochem. Soc. Trans. *41*, 1392–1400.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol *13*, 722–736.

Marraffini, L.A., and Sontheimer, E.J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature *463*, 568–571.

McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Mol Cell *64*, 616–623.

Mekler, V., Minakhin, L., and Severinov, K. (2017). Mechanism of duplex DNA destabilization by RNA-guided Cas9 nuclease during target interrogation. Proc. Natl. Acad. Sci. *114*, 5443–5448.

Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T., and Lenski, R.E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. Science *335*, 428–432.

Modell, J.W., Jiang, W., and Marraffini, L.A. (2017). CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *544*, 101–104.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol *60*, 174–182.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mulepati, S., and Bailey, S. (2011). Structural and biochemical analysis of the nuclease domain of the clustered regularly interspaced short palindromic repeat (CRISPR) associated protein 3(CAS3). J. Biol. Chem.

Nam, K.H., Kurinov, I., and Ke, A. (2011). Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA binding activity. J. Biol. Chem. *286*, 30759–30768.

Nam, K.H., Ding, F., Haitjema, C., Huang, Q., DeLisa, M.P., and Ke, A. (2012). Double-stranded endonuclease activity in Bacillus halodurans clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. J. Biol. Chem. *287*, 35943–35952.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. Cell *156*, 935–949.

Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., and Nureki, O. (2015). Crystal Structure of Staphylococcus aureus Cas9. Cell *162*, 1113–1126.

Nozawa, T., Furukawa, N., Aikawa, C., Watanabe, T., Haobam, B., Kurokawa, K., Maruyama, F., and Nakagawa, I. (2011). CRISPR inhibition of prophage acquisition in Streptococcus pyogenes. PLoS One *6*, e19543.

Nunez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat. Struct. Mol. Biol.

Nunez, J.K., Lee, A.S., Engelman, A., and Doudna, J.A. (2015a). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature *519*, 193–198.

Nunez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N., and Doudna, J.A. (2015b). Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature *527*, 535–538.

Nunez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., and Doudna, J.A. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol Cell *62*, 824–833.

Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D.K., and Aravin, A.A. (2013). Bacterial Argonaute Samples the Transcriptome to Identify Foreign DNA. Mol. Cell *51*, 594–605.

Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K., Stahl, B., Barrangou, R., and Banfield, J.F. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nat Commun *4*, 1430.

Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B.C., Barrangou, R., and Banfield, J.F. (2015). CRISPR Immunity Drives Rapid Phage Genome Evolution in *Streptococcus thermophilus*. MBio *6*.

Pawluk, A., Davidson, A.R., and Maxwell, K.L. (2017). Anti-CRISPR: discovery, mechanism and function. Nat. Rev. Microbiol. *16*, 12–17.

Peng, R., Xu, Y., Zhu, T., Li, N., Qi, J., Chai, Y., Wu, M., Zhang, X., Shi, Y., Wang, P., et al. (2017). Alternate binding modes of anti-CRISPR viral suppressors AcrF1/2 to Csy surveillance complex revealed by cryo-EM structures. Cell Res. *27*, 853–864.

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology *151*, 653–663.

Pyenson, N.C., Gayvert, K., Varble, A., Elemento, O., and Marraffini, L.A. (2017). Broad Targeting Specificity during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. Cell Host Microbe *22*, 343-353.e3.

Quiberoni, A., Moineau, S., Rousseau, G.M., Reinheimer, J., and Ackermann, H.-W. (2010). Streptococcus thermophilus bacteriophages. Int. Dairy J. *20*, 657–664.

Radovčić, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E.L., and Ivančić-Baće, I. (2018). CRISPR–Cas adaptation in Escherichia coli requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5′ ssDNA exonucleases. Nucleic Acids Res.

Rain, J.-C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., et al. (2001). The protein–protein interaction map of Helicobacter pylori. Nature *409*, 211–215.

Rauch, B.J., Silvis, M.R., Hultquist, J.F., Waters, C.S., McGregor, M.J., Krogan, N.J., and Bondy-Denomy, J. (2017). Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. Cell *168*, 150-158.e10.

Redding, S., Sternberg, S.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell *163*, 854–865.

Richter, C., Gristwood, T., Clulow, J.S., and Fineran, P.C. (2012). In vivo protein interactions and complex formation in the Pectobacterium atrosepticum subtype I-F CRISPR/Cas System. PLoS One *7*, e49549.

Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H., and Fineran, P.C. (2014). Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res *42*, 8516–8526.

Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. *42*, W320–W324.

Rollie, C., Graham, S., Rouillon, C., and White, M.F. (2018). Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Res. *46*, 1007–1020.

Rollins, M.F., Chowdhury, S., Carter, J., Golden, S.M., Wilkinson, R.A., Bondy-Denomy, J., Lander, G.C., and Wiedenheft, B. (2017). Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proc. Natl. Acad. Sci. 201616395.

Samai, P., Smith, P., and Shuman, S. (2010). Structure of a CRISPR-associated protein Cas2 from *Desulfovibrio vulgaris*. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. *66*, 1552–1556.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (NY: Cold Spring Harbor Laboratory Press).

Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A., and Severinov, K. (2013). High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA Biol *10*, 716–725.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature *494*, 489–491.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U A *108*, 10098–10103.

Shiimori, M., Garrett, S.C., Chambers, D.P., Glover, C.V.C., Graveley, B.R., and Terns, M.P. (2017). Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in Pyrococcus furiosus. Nucleic Acids Res. *45*, 11281–11294.

Shiimori, M., Garrett, S.C., Graveley, B.R., and Terns, M.P. (2018). Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. Mol. Cell *70*, 814-824.e6.

Shin, J., Jiang, F., Liu, J.-J., Bray, N.L., Rauch, B.J., Baik, S.H., Nogales, E., Bondy-Denomy, J., Corn, J.E., and Doudna, J.A. (2017). Disabling Cas9 by an anti-CRISPR DNA mimic. Sci. Adv. *3*, e1701620.

Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2016). Molecular recordings by directed CRISPR spacer acquisition. Science *353*, aaf1175.

Shmakov, S., Savitskaya, E., Semenova, E., Logacheva, M.D., Datsenko, K.A., and Severinov, K. (2014). Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res *42*, 5907–5916.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell *60*, 385–397.

Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A.M., and Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science *351*, aad4234.

Singh, P., Panchaud, A., and Goodlett, D.R. (2010). Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. Anal. Chem. *82*, 2636–2642.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J. *30*, 1335–1342.

Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. EMBO J. *32*, 385–394.

Siringan, P., Connerton, P.L., Cummings, N.J., and Connerton, I.F. (2014). Alternative bacteriophage life cycles: the carrier state of Campylobacter jejuni. Open Biol. *4*, 130200–130200.

Smith, G.R. (2012). How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev *76*, 217–228.

Staals, R.H., Jackson, S.A., Biswas, A., Brouns, S.J., Brown, C.M., and Fineran, P.C. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nat Commun *7*, 12853.

Stasi, M., De Luca, M., and Bucci, C. (2015). Two-hybrid-based systems: Powerful tools for investigation of membrane traffic machineries. J. Biotechnol. *202*, 105–117.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Swarts, D.C., Mosterd, C., van Passel, M.W., and Brouns, S.J. (2012). CRISPR interference directs strand specific spacer acquisition. PLoS One *7*, e35888.

Swarts, D.C., Jore, M.M., Westra, E.R., Zhu, Y., Janssen, J.H., Snijders, A.P., Wang, Y., Patel, D.J., Berenguer, J., Brouns, S.J.J., et al. (2014). DNA-guided DNA interference by a prokaryotic Argonaute. Nature *507*, 258–261.

Swarts, D.C., Hegge, J.W., Hinojo, I., Shiimori, M., Ellis, M.A., Dumrongkulraksa, J., Terns, R.M., Terns, M.P., and van der Oost, J. (2015). Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA. Nucleic Acids Res. *43*, 5120–5129.

Swarts, D.C., Szczepaniak, M., Sheng, G., Chandradoss, S.D., Zhu, Y., Timmers, E.M., Zhang, Y., Zhao, H., Lou, J., Wang, Y., et al. (2017). Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. Mol. Cell *65*, 985-998.e6.

Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. *111*, 9798–9803.

Truglio, J.J., Croteau, D.L., Van Houten, B., and Kisker, C. (2006). Prokaryotic Nucleotide Excision Repair: The UvrABC System. Chem. Rev. *106*, 233–252.

Van Houten, B., Croteau, D.L., DellaVecchia, M.J., Wang, H., and Kisker, C. (2005). 'Close-fitting sleeves': DNA damage recognition by the UvrABC nuclease system. Mutat. Res. Mol. Mech. Mutagen. *577*, 92–117.

Van Orden, M.J., Klein, P., Babu, K., Najar, F.Z., and Rajan, R. (2017). Conserved DNA motifs in the type II-A CRISPR leader region. PeerJ *5*, e3161.

Vogan, A.A., and Higgs, P.G. (2011). The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. Biol. Direct *6*, 1.

Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K., Logacheva, M., Wiedenheft, B., Davidson, A.R., Severinov, K., et al. (2015). Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res *43*, 10848–10860.

Walker, M.J., Barnett, T.C., McArthur, J.D., Cole, J.N., Gillen, C.M., Henningham, A., Sriprakash, K.S., Sanderson-Smith, M.L., and Nizet, V. (2014). Disease Manifestations and Pathogenic Mechanisms of Group A Streptococcus. Clin. Microbiol. Rev. *27*, 264–301.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., and Wang, Y. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

Wei, Y., Terns, R.M., and Terns, M.P. (2015a). Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev *29*, 356–361.

Wei, Y., Chesne, M.T., Terns, R.M., and Terns, M.P. (2015b). Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Res *43*, 1749–1758.

Westermarck, J., Ivaska, J., and Corthals, G.L. (2013). Identification of Protein Interactions Involved in Cellular Signaling. Mol. Cell. Proteomics *12*, 1752–1763.

Westra, E.R., and Brouns, S.J.J. (2012). The rise and fall of CRISPRs - dynamics of spacer acquisition and loss: Dynamics of CRISPR adaptation. Mol. Microbiol. *85*, 1021–1025.

Westra, E.R., van Erp, P.B., Künne, T., Wong, S.P., Staals, R.H., Seegers, C.L., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., et al. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell *46*, 595–605.

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., and Doudna, J.A. (2009). Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure *17*, 904–912.

Wigley, D.B. (2013). Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. Nat Rev Microbiol *11*, 9–13.

Williams, E., Lowe, T.M., Savas, J., and DiRuggiero, J. (2007). Microarray analysis of the hyperthermophilic archaeon Pyrococcus furiosus exposed to gamma irradiation. Extremophiles *11*, 19–29.

Wright, A.V., and Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. Nat. Struct. Mol. Biol. *23*, 876–883.

Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E., and Doudna, J.A. (2017). Structures of the CRISPR genome integration complex. Science *357*, 1113–1118.

Xiao, Y., Ng, S., Hyun Nam, K., and Ke, A. (2017). How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. Nature *550*, 137–141.

Xue, C., Seetharam, A.S., Musharova, O., Severinov, K., J. Brouns, S.J., Severin, A.J., and Sashital, D.G. (2015). CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Res. *43*, 10831–10847.

Xue, C., Whitis, N.R., and Sashital, D.G. (2016). Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Mol. Cell *64*, 826–834.

Yamada, M., Watanabe, Y., Gootenberg, J.S., Hirano, H., Ran, F.A., Nakane, T., Ishitani, R., Zhang, F., Nishimasu, H., and Nureki, O. (2017). Crystal Structure of the Minimal Cas9 from Campylobacter jejuni Reveals the Molecular Diversity in the CRISPR-Cas9 Systems. Mol. Cell *65*, 1109-1121.e3.

Yang, H., and Patel, D.J. (2017). Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9. Mol. Cell *67*, 117-127.e5.

Yoganand, K.N.R., Sivathanu, R., Nimkar, S., and Anand, B. (2017). Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic Acids Res. *45*, 367–381.

Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res *40*, 5569–5576.

Zander, A., Holzmeister, P., Klose, D., Tinnefeld, P., and Grohmann, D. (2014). Single-molecule FRET supports the two-state model of Argonaute action. RNA Biol. *11*, 45–56.

Zander, A., Willkomm, S., Ofer, S., van Wolferen, M., Egert, L., Buchmeier, S., Stöckl, S., Tinnefeld, P., Schneider, S., Klingl, A., et al. (2017). Guide-independent DNA cleavage by archaeal Argonaute from Methanocaldococcus jannaschii. Nat. Microbiol. *2*, 17034.

Zhang, J., Kasciukovic, T., and White, M.F. (2012). The CRISPR associated protein Cas4 Is a 5’ to 3’ DNA exonuclease with an iron-sulfur cluster. PLoS One *7*, e47232.

11 Acknowledgements

The conception of this thesis would not have been possible without all the supportive people around me. I would like to take this opportunity and thank everyone for their contributions.

I would like to thank my PhD supervisor, Emmanuelle Charpentier, for giving me the chance and complete freedom to work on this interesting PhD topic, in a work environment which is outstanding in many ways. Thank you for giving me the opportunity to do my PhD research across two countries, three cities and multiple institutes.

I would like to express my gratitude to my PhD committee members, Matthew Francis, Felipe Cava and Caroline Grabbe, from Umeå University, for their helpful discussion and input to my PhD projects. Thank you Kürşad Turgay for reading my PhD thesis manuscript and providing helpful comments.

A big Thank You to the administrative staffs in MiMS, HZI and MPIIB for supporting me with all the administrative work and translation, especially Maria, Helga, Lya and Lisa.

Yan Yan, Katja, Sandra Augustin and Annette, I am very grateful for your assistance inside and also outside the lab. Thank you Yan Yan for helping me to paint my living room when I moved to Berlin. Special thanks to Katja for her help in my experiments, German questions and the fun conversations. I would like to thank my intern student Ann Kathrin for her experimental help.

Special thanks to Anaïs, who gave me a lot of scientific help, especially at the beginning of my PhD. Thanks for the fun moments and making wonderful galette. Ines, I would like to show my appreciation to your scientific support and personal help. Anne-Laure, thanks for your kind suggestions to my PhD projects, hospitality when I went back to Umeå, teaching me how to bake macarons, and the very important French vocabulary. Geetanjali, thanks for your scientific help, encouragement to do art and the fun time outside the lab. Khan, for being a very helpful colleague. Special thanks to Majda, who is always supportive and helpful to my scientific work, especially giving me a lot of input on my thesis. Majda, thanks for also being such a nice and sincere friend outside the lab and sharing all the happy and sad moments. I would also like to thank Thibaud, Frank and Hagen, for their scientific help. Thank you Frank for reading my thesis.

Thanks a lot to all the former and current members in MiMS, HZI and MPIIB for helping and sharing, as well as wonderful time and great experience. Thanks Andrés, James, Laura, Lina, Dior, Sandra Franch-Arroyo, Vanessa, Stefan, Sarah and Thomas, for all the help in the lab and the good time outside the lab. Thank you Marléne for giving me a big hug when I was upset.

I would like to express my gratitude to my friends, who always support and encourage me: Kaiyi, Martin, Loo Wee, Suin, Poh Shin, Sook Ting, Pik Foong, Jun Han, Wai Siang, Meow Lin, Khek Chian, Chantal, Anandi, Mridula, Swetlana, Sandrine, Carrie, Eeelee and Shook Lin.

Thousands of thanks to Matthias, who has always been there for me during all the ups and downs, for being understanding and supportive. Thank you for believing in me. I would not have gone so far without your support.

And last, but not least, I would like thank my mother for her constant support, understanding and blessings. Thanks to my brothers, Chee Leong and Chee Yien, for taking care of our parents for all these years when I have been abroad. I would like express my gratitude to my late father — I hope that he could see the completion of this thesis from heaven.